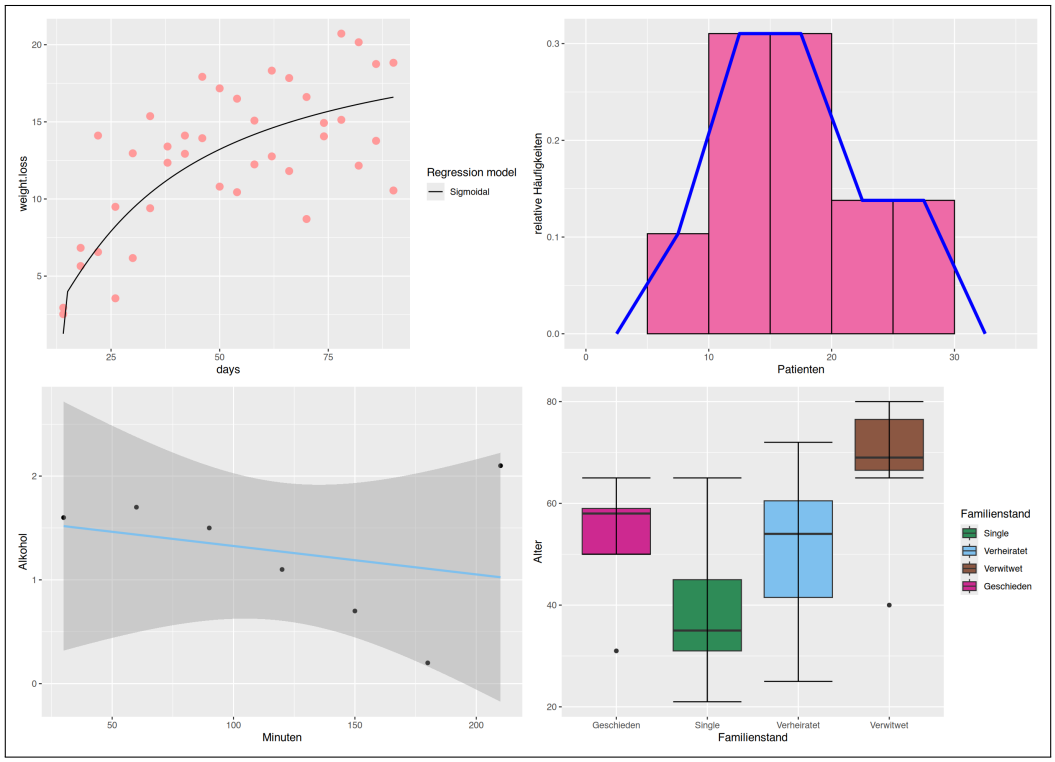
Angewandte Übungen in R



Über 300 Aufgaben und Lösungswege für 80 Fragen in 12 Themengebieten der biomedizinischen Statistik.

zusammengestellt von
Prof. Dr. Jörg große Schlarmann

Version vom 25.06.2024

Inhaltsverzeichnis

Angewandte Übungen in R

In diesem Script wollen wir Übungsaufgaben zu verschiedenen Teilen der Statistik vorstellen und lösen. Die Aufgaben stammen von Gimeno et al. (2022)[1]. Dort werden die Lösungswege unter Verwendung der Software RKWard[2] besprochen. Im vorliegenden Script sollen die Lösungen "zu Fuß" erarbeitet werden. Hierbei kann das freie Nachschlagewerk von große Schlarmann (2024) hilfreich sein.

Lizenz

Zitationsvorschlag

große Schlarmann, J (2024): "Angewandte Übungen in R", Hochschule Niederrhein, https://www.produnis.de/R/exercise.html

```
@book{grSchl_exeRcise,
    author = {{große Schlarmann}, Jörg},
    title = {Angewandte Übungen in R},
    year = {2024},
    publisher = {Hochschule Niederrhein},
    address = {Krefeld},
    copyright = {CC BY-NC-SA 4.0},
    url = {https://www.produnis.de/R/exercise.html},
    language = {de},
}
```

[1] siehe https://github.com/asalber/statistics_practice_rkteaching
[2] siehe https://rkward.kde.org/
[3] siehe https://creativecommons.org/licenses/by-nc-sa/4.0/

1 Aufgabenstellung

Versuchen Sie zunächst selbst eine Lösung zu finden bevor Sie sich die Auflösungen anschauen.

1.1 Häufigkeitsverteilungen

1.1.1 Aufgabe 1.1.1 Kinder in Familien

Für 25 Familien liegt die Anzahl an Kindern vor:

1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2

a) Erstellen Sie ein Datenframe mit der Variable `Kinder` und übertragen Sie die Daten.
b) Erzeugen Sie eine einfache Häufigkeitstabelle
c) Erzeugen Sie ein Balkendiagramm der Häufigkeiten
d) Erzeugen Sie eine vollständige Häufigkeitstabelle, inklusive absoluter, relativer und jeweils kumulativer Häufigkeiten

Für die Lösung siehe Abschnitt 2.1

1.1.2 Aufgabe 1.1.2 Patienten in der Notaufnahme

Den gesamten November über wurde die Anzahl an Patienten in der Notaufnahme erhoben

15	23	12	10	28	50	12	17	20	21	18	13	11	12	26
30	6	16	19	22	14	17	21	28	9	16	13	11	16	20

a) Erstellen Sie ein Datenframe mit der Variable `Patienten` und übertragen Sie die Daten.
b) Erzeugen Sie ein Boxplot. Gibt es Ausreißer? Wenn ja, entfernen Sie diese, bevor Sie weitermachen.
c) Erzeugen Sie eine Häufigkeitstabelle, welche die Daten in 5 Klassen gruppiert.
d) Erzeugen Sie ein Histogram der klassierten absoluten Häufigkeiten.
e) Erzeugen Sie ebenso Histogramme der relativen und jeweils kumulativen Häufigkeiten, inklusive Polygonzügen.

Für die Lösung siehe Abschnitt 2.2

1.1.3 Aufgabe 1.1.3 Blutgruppen

Von 30 Personen wurden die Blutgruppen wie folgt bestimmt:

A, B, B, A, AB, 0, 0, A, B, B, A, A, A, A, AB,
A, A, A, B, 0, B, B, B, A, A, A, 0, A, AB, 0.

a) Erstellen Sie ein Datenframe mit der Variable `Blutgruppe` und übertragen Sie die Daten.
b) Erzeugen Sie eine Häufigkeitstabelle
c) Erzeugen Sie ein Kreisdiagramm

Für die Lösung siehe Abschnitt 2.3

1.1.4 Aufgabe 1.1.4 Familienstand

Das Alter und der Familienstand von 28 Personen wurden wie folgt erhoben:

Familienstand	Alter								
Single	31	45	35	65	21	38	62	22	31
Verheiratet	72	39	62	59	25	44	54		
Verwitwet	80	68	65	40	78	69	75		
Geschieden	31	65	59	58	50				

a) Erstellen Sie ein Datenframe mit den Variablen `Alter` und `Familienstand` und übertragen Sie die Daten.
b) Erzeugen Sie für jeden `Familienstand` eine Häufigkeitstabelle des `Alters`.
c) Erzeugen Sie für jeden `Familienstand` eine Boxplot des `Alters`. Gibt es Ausreißer? In welcher Gruppe streut das Alter am meisten?
d) Erzeugen Sie für jeden `Familienstand` eine Histogram des `Alters`. Wie unterscheiden sich die Histogramme?

Für die Lösung siehe Abschnitt 2.4

1.1.5 Aufgabe 1.1.5 Handballverletzungen

Die Anzahl der Verletzungen von Handballspielern eines Teams wurden wie folgt erhoben:

0, 1, 2, 1, 3, 0, 1, 0, 1, 2, 0, 1, 1, 1, 2, 0, 1, 3, 2, 1, 2, 1, 0, 1

a) Erstellen Sie eine Häufigkeitstabelle

b) Erzeugen Sie ein Säulendiagramm der relativen und kumulativen relativen Häufigkeiten.
c) Erzeugen Sie ein Boxplot

Für die Lösung siehe Abschnitt 2.5

1.1.6 Aufgabe 1.1.6 Körpergröße

Von 30 Studierenden wurde die Körpergröße gemessen

179, 173, 181, 170, 158, 174, 172, 166, 194, 185,
162, 187, 198, 177, 178, 165, 154, 188, 166, 171,
175, 182, 167, 169, 172, 186, 172, 176, 168, 187

a) Erstellen Sie ein Histogram der Körpergröße mit Klassen von 150cm bis 200cm, die jeweils 10cm breit sind.
b) Gibt es Ausreißer?

Für die Lösung siehe Abschnitt 2.6

1.1.7 Aufgabe 1.1.7 Neugeborene

Der Datensatz `neonates` von rk.Teaching[4] enthält Informationen über eine Stichprobe von 320 Neugeborenen, die im Laufe eines Jahres nach normaler Schwangerschaftsdauer geboren wurden.

a) Erstellen Sie die Häufigkeitstabelle des APGAR-Scores nach 1 Minute. Wenn ein Score von 3 oder weniger anzeigt, dass das Neugeborene in einem kritischen Zusatand ist, wie viel Prozent der Neugeborenen in der Stichprobe sind dann in einem kritischen Zustand?

b) Erstellen Sie die Häufigkeitstabelle des Geburtsgewichts der Neugeborenen, indem Sie die Daten in Klassen mit einer Breite von 0,5 kg von 2 bis 4,5 kg einteilen. Welches Intervall enthält die meisten Neugeborenen?

c) Vergleichen Sie die Häufigkeitsverteilung des APGAR-Scores nach 1 Minute für Mütter unter 20 Jahren und für Mütter über 20 Jahren. Welche Gruppe hat mehr Neugeborene in kritischem Zustand?

d) Vergleichen Sie die relative Häufigkeitsverteilung des Geburtsgewichts der Neugeborenen, je nachdem, ob die Mutter während der Schwangerschaft geraucht hat oder nicht. Wenn ein Gewicht unter 2,5 kg als niedriges Gewicht gilt, welche Gruppe hat einen höheren Prozentsatz an Neugeborenen mit niedrigem Gewicht?

e) Berechnen Sie die Prävalenz von Neugeborenen mit niedrigem Gewicht für Mütter, die vor der Schwangerschaft geraucht haben, und den Nichtraucherinnen.

f) Berechnen Sie das relative Risiko eines niedrigen Geburtsgewichts des Neugeborenen, wenn die Mutter während der Schwangerschaft raucht, im Vergleich dazu, wenn die Mutter nicht raucht.

g) Erstellen Sie ein Balkendiagramm des APGAR-Scores nach 1 Minute. Welcher Score ist am häufigsten?

h) Erstellen Sie das Balkendiagramm der kumulierten relativen Häufigkeit des APGAR-Scores nach 1 Minute. Unter welchem Wert liegen die Hälfte der Neugeborenen?

i) Vergleichen Sie die Balkendiagramme der relativen Häufigkeitsverteilungen des APGAR-Scores nach 1 Minute, je nachdem, ob die Mutter während der Schwangerschaft geraucht hat oder nicht. Welche Schlussfolgerungen können gezogen werden?

j) Erstellen Sie ein Histogramm der Geburtsgewichte der Neugeborenen mit Klassenbreiten von 0,5 kg von 2 bis 4,5 kg. Welche Klasse enthält die meisten Neugeborenen?

k) Vergleichen Sie die relativen Häufigkeitshistogramme der Geburtsgewichte der Neugeborenen, mit Klassenbreiten von 0,5 kg von 2 bis 4,5 kg, je nachdem, ob die Mutter während der Schwangerschaft geraucht hat oder nicht. Welche Gruppe hat Neugeborene mit geringeren Gewichten?

l) Vergleichen Sie die relativen Häufigkeitshistogramme der Geburtsgewichte der Neugeborenen, mit Klassenbreiten von 0,5 kg von 2 bis 4,5 kg, je nachdem, ob die Mutter vor der Schwangerschaft geraucht hat oder nicht. Welche Schlussfolgerungen können gezogen werden?

m) Erstellen Sie ein Boxplot der Geburtsgewichte der Neugeborenen. Welcher Gewichtsbereich kann in der Stichprobe als normal angesehen werden? Gibt es Ausreißer in der Stichprobe?

n) Vergleichen Sie die Boxplots der Geburtsgewichte der Neugeborenen je nachdem, ob die Mutter während der Schwangerschaft geraucht hat oder nicht und ob die Mutter unter 20 oder über 20 Jahre alt war. Welche Gruppe hat eine größere zentrale Streuung? Welche Gruppe hat Neugeborene mit geringerem Gewicht?

o) Vergleichen Sie die Boxplots der APGAR-Scores nach 1 Minute und nach 5 Minuten. Welche Variable hat eine größere zentrale Streuung?

Für die Lösung siehe Abschnitt 2.7

[4] https://github.com/rkward-community/rk.Teaching, auch verfügbar unter https://www.produnis.de/R/data/neonates.RData

1.2 Stichprobenstatistik

Bei diesen Aufgaben geht es vor allem um Lage- und Streuungskenngrößen.

1.2.1 Aufgabe 1.2.1 Kinder in Familien

Die Anzahl an Kindern in einer Stichprobe aus 25 Familien sind:

1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2

a) Erstellen Sie ein Datenframe mit der Variable `Kinder` und übertragen Sie die Daten.
b) Berechnen Sie das arithmetische Mittel, die Varianz sowie die Standardabweichung für die Anzahl an Kindern.
c) Berechnen Sie die Quartile, die Spannweite, den Interquartilsabstand, das dritte Dezil sowie das 68te Perzentil.

Für die Lösung siehe Abschnitt 2.8

1.2.2 Aufgabe 1.2.2 Patienten in Notaufnahme

Den gesamten November über wurde die Anzahl an Patienten in der Notaufnahme erhoben

15	23	12	10	28	50	12	17	20	21	18	13	11	12	26
30	6	16	19	22	14	17	21	28	9	16	13	11	16	20

a) Erstellen Sie ein Datenframe mit der Variable `Patienten` und übertragen Sie die Daten.
b) Berechnen Sie das arithmetische Mittel, die Varianz, die Standardabweichung und den Variationskoeffizienten.
c) Berechnen Sie die Skewness (Schiefe) und Kurtosis (“Spitzigkeit”) und interpretieren Sie die Werte.

Für die Lösung siehe Abschnitt 2.9

1.2.3 Aufgabe 1.2.3 Studierendenbewertung

Im letzten R-Kurs haben 20 Studierende folgende Abschlussbewertungen erhalten

SS, AP, SS, AP, AP, NT, NT, AP, SB, SS
SB, SS, AP, AP, NT, AP, SS, NT, SS, NT

a) Erstellen Sie ein Datenframe mit der Variable `Bewertung` und übertragen Sie die Daten.

b) Wandeln Sie die `Bewertung` in Punkte um, nach dem Schema "`SS`" = 2,5 | "`AP`" = 6 | "`NT`" = 8 | "`SB`" = 9,5.
c) Bestimmen Sie den Median und den Interquartilsabstand.

Für die Lösung siehe Abschnitt 2.10

1.2.4 Aufgabe 1.2.4 Körpergröße nach Geschlecht

Von 30 Studierenden wurde die Körpergröße wie folgt gemessen:

Geschlecht	Größe
weiblich	173, 158, 174, 166, 162, 177, 165, 154, 166, 182, 169, 172, 170, 168
männlich	179, 181, 172, 194, 185, 187, 198, 178, 188, 171, 175, 167, 186, 172, 176, 187

a) Erstellen Sie ein Datenframe mit den Variable `Geschlecht` und `Koerpergroesse` und übertragen Sie die Daten.
b) Bestimmen Sie in Abhängigkeit zum `Geschlecht` das arithmetische Mittel, den Median, die Varianz, die Standardabweichung sowie die Quartile.

Für die Lösung siehe Abschnitt 2.11

1.2.5 Aufgabe 1.2.5 Handballverletzungen

Die Anzahl der Verletzungen von Handballspielern eines Teams wurden wie folgt erhoben:

0, 1, 2, 1, 3, 0, 1, 0, 1, 2, 0, 1, 1, 1, 2, 0, 1, 3, 2, 1, 2, 1, 0, 1

a) Bestimmen Sie das arithmetische Mittel, den Median, die Varianz sowie die Standardabweichung der Verletzungen.
b) Bestimmen Sie die Skewness und Kortosis der Verteilung.
c) Berechnen Sie das vierte und achte Dezil der Verteilung.

Für die Lösung siehe Abschnitt 2.12

1.2.6 Aufgabe 1.2.6 Blutdruckmessung

Wir möchten die Zuverlässigkeit zweier Blutdruckmonitore bestimmen. Gerät 1 misst den Blutdruck am Handgelenk, Gerät 2 am Unterarm. Es wurden 8 Messungen mit jedem Gerät bei der selben Person durchgeführt, wobei folgende systolischen Werte gemessen wurden:

Position	Messdaten
Unterarm	111, 109, 112, 111, 113, 113, 114, 111
Handgelenk	115, 113, 117, 116, 112, 112, 117, 112

Welcher Monitor funktioniert besser?

Für die Lösung siehe Abschnitt 2.13

1.2.7 Aufgabe 1.2.7 Alter und Familienstand

Das Alter und der Familienstand von 28 Personen wurden wie folgt erhoben:

Familienstand	Alter								
Single	31	45	35	65	21	38	62	22	31
Verheiratet	72	39	62	59	25	44	54		
Verwitwet	80	68	65	40	78	69	75		
Geschieden	31	65	59	58	50				

a) Bestimmen Sie das arithmetische Mittel, den Median, die Varianz sowie die Standardabweichung des `Alters` für jeden `Familienstand`.
b) Welche Gruppe hat den "besten" Mittelwert?

Für die Lösung siehe Abschnitt 2.14

1.2.8 Aufgabe 1.2.8 Tabak, Alkohol und Blutdruck

Eine Studie möchte den möglichen Zusammenhang zwischen dem Blutdruck und dem Alkohol- und Tabakkonsum untersuchen. Hierzu wurden folgende Daten von 25 Personen erhoben.

Rauchen	ja	nein	ja	ja	ja	nein	nein	ja	nein	ja	nein	ja	nein
Alkohol	nein	nein	ja	ja	nein	nein	ja	ja	nein	ja	nein	ja	ja
Blutdruck	80	92	75	56	89	93	101	67	89	63	98	58	91

Rauchen	ja	nein	nein	ja	nein	nein	nein	ja	nein	ja	nein	ja
Alkohol	ja	nein	ja	ja	nein	nein	ja	ja	ja	nein	ja	nein
Blutdruck	71	52	98	104	57	89	70	93	69	82	70	49

a) Vergleichen Sie das arithmetische Mittel, die Standardabweichung, die Skewness und Kurtosis des `Blutdrucks` zwischen Rauchern und Nichtrauchern.
b) Vergleichen Sie die selben Werte zwischen der Alkohol- und Nicht-Alkoholgruppe.
c) Vergleichen Sie die selben Werte zwischen der Raucher- und Alkoholgruppe, zwischen der Raucher- und Nicht-Alkoholgruppe, der Nichtraucher- und Alkoholgruppe sowie der Nichtraucher- und Nicht-Alkoholgruppe.

Für die Lösung siehe Abschnitt 2.14

1.3 Lineare Regression

1.3.1 Aufgabe 1.3.1 X und Y

Bei 10 Personen wurden `x` und `y` erhoben.

x	0	1	2	3	4	5	6	7	8	9
y	2	5	8	11	14	17	20	23	26	29

a) Erstellen Sie ein Datenframe mit den Variablen `x` und `y`.
b) Erzeugen Sie ein Scatterplot von `x` und `y`. Bestimmen Sie anhand des Plots, welche Regressionsfunktion die Daten am besten erklären würde.
c) Führen Sie die Regression durch.
d) Fügen Sie die Regressionsfunktion `y erklärt durch x` dem Plot hinzu.
e) Fügen Sie die Regressionsfunktion `x erklärt durch y` ebenfalls dem Plot hinzu, aber in roter Farbe.
f) Wie groß sind die Residuen?

Für die Lösung siehe Abschnitt 2.16

1.3.2 Aufgabe 1.3.2 Lernen und Durchfallen

Eine Studie gibt vor, den Zusammenhang zwischen den täglichen Lernstunden und der Anzahl nicht bestandener Prüfungen im Semester zu untersuchen. Bei 30 Studierenden wurden folgende Werte erhoben:

Lernzeit	durchgefallen	Lernzeit	durchgefallen	Lernzeit	durchgefallen
3.5	1	2.2	2	1.3	4
0.6	5	3.3	0	3.1	0
2.8	1	1.7	3	2.3	2
2.5	3	1.1	3	3.2	2
2.6	1	2.0	3	0.9	4
3.9	0	3.5	0	1.7	2
1.5	3	2.1	2	0.2	5
0.7	3	1.8	2	2.9	1
3.6	1	1.1	4	1.0	3
3.7	1	0.7	4	2.3	2

a) Erstellen Sie ein Datenframe mit den Variablen `Lernen` und `Durchgefallen`.
b) Erzeugen Sie eine Kreuztabelle der Variablen `Lernen` und `Durchgefallen`.
c) Führen Sie eine lineare Regression `Durchgefallen erklärt durch Lernen` durch und plotten Sie Ihr Ergebnis.
d) Wie lauten die Regressionskoeffizient des Modells, und wie ist er zu interpretieren?
e) Ist das soeben erstellte Modell *besser* als das in Abschnitt 1.3.1 berechnete? Vergleichen Sie zur Beantwortung die Residuen beider Modelle.
f) Berechnen Sie den linearen Bestimmungskoeffizient und den Korrelationskoeffizient. Ist das lineare Modell ein gutes Modell, um die Beziehung zwischen den gescheiterten Prüfungen und den täglichen Studienzeiten zu erklären? Wie viel Prozent der Variabilität der durchgefallenen Prüfungen wird durch das lineare Modell erklärt?
g) Benutzen Sie das lineare Modell, um die Anzahl an durchgefallenen Prüfungen für einen Studierenden zu bestimmen, der 3 Stunden Lernzeit investiert hat. Wie glaubwürdig ist die Vorhersage?
h) Wie viele Stunden Lernzeit wird benötigt, um alle Kurse zu bestehen?

Für die Lösung siehe Abschnitt 2.17

1.3.3 Aufgabe 1.3.3 Metabolismus

Um herauszufinden, wie der Körper Alkohol verstoffwechselt, hat ein Proband einen Liter Wein zügig getrunken. Anschließend wurde alle 30 Minuten der Blutalkoholspiegel gemessen.

Minuten	30	60	90	120	150	180	210
Alkohol (g/l)	1.6	1.7	1.5	1.1	0.7	0.2	2.1

a) Erstellen Sie ein Datenframe mit den Variablen `Minuten` und `Alkohol`.
b) Bestimmen Sie den passenden Korrelationskoeffizienten. Werden die Daten ausreichend gut durch das Modell beschrieben?
c) Plotten Sie das lineare Regressionsmodell `Alkohol erklärt durch Minuten`. Gibt es Punkte mit großen Residuen? Wenn ja, entfernen Sie diese und führen die Berechnungen erneut durch. Hat sich der Korrelationskoeffizient verbessert?
d) Mit welcher Geschwindigkeit wird der Alkohol pro Minute verstoffwechselt?
e) Wenn es gesetzlich erlaubt wäre, mit einem Blutalkoholwert von $0,3$ g/l Auto zu fahren, wie lange muss die Person warten, nachdem sie 1 Liter Weingetrunken hat, um wieder fahrtüchtig zu sein? Wie zuverlässig ist diese Vorhersage?

Für die Lösung siehe Abschnitt 2.18

1.3.4 Aufgabe 1.3.4 Alter und Körpergröße

Im Datensatz `age.height` von rk.Teaching[5] sind Alter und Körpergröße von 30 Probanden enthalten.

a) Laden Sie den Datensatz `age.height` in Ihre R-Session.
b) Berechnen Sie die Regressionsgerade `Größe erklärt durch Alter`. Ist das lineare Modell geeignet, den Zusammenhang zwischen `Alter` und `Körpergröße` zu erklären?
c) Erstellen Sie eine Punktwolke inklusive der Regressionsgeraden. Ab welchem Alter ändert sich die Punktetendenz?
d) Erstellen Sie eine Gruppierungsvariable, welche `Alter` in einen ordinalen Faktor mit den Ausprägungen "`jünger als 20`" und "`20 und älter`" einteilt.
e) Führen Sie die lineare Regressionsanalyse für beide Gruppen erneut durch. In welcher Gruppe wird der Zusammenhang zwischen `Alter` und `Körpergröße` am besten erklärt?
f) Plotten Sie die Modelle.
g) Welche Körpergröße sagt Ihr Modell für eine 14jährige Person vorher, und welche für eine 38jährige Person?

Für die Lösung siehe Abschnitt 2.19

[5] https://github.com/rkward-community/rk.Teaching, auch verfügbar unter https://www.produnis.de/R/data/age.height.RData

1.3.5 Aufgabe 1.3.5 Wirksamkeitsverlust

Eine Studie Untersucht den Wirksamkeitsverlust eines Medikaments, das über Jahre von vielen Probanden eingenommen wurde. Folgende Aussagen zur Wirksamkeit konnten aus den Daten ermittelt werden.

Jahr	1	2	3	4	5
Wirksamkeit (%)	96	84	70	58	52

a) Führen Sie eine lineare Regression `Wirksamkeit erklärt durch Jahr` durch und plotten Sie Ihr Ergebnis.
b) Wie große ist der jährliche Wirksamkeitsverlust in %?
c) Nach wie vielen Jahren ist die Wirksamkeit bei 80%, und nach wie vielen bei 0%? Sind beide Werte gleich zuverlässig?

Für die Lösung siehe Abschnitt 2.20

1.3.6 Aufgabe 1.3.6 Dosierung

In einer Studie über die Wirkung verschiedener Dosen eines Medikaments erhielten 2 Patienten 2 mg und benötigten 5 Tage zur Heilung, 4 Patienten erhielten 2 mg und benötigten 6 Tage zur Heilung, 2 Patienten erhielten 3 mg und benötigten 3 Tage zur Heilung, 4 Patienten erhielten 3 mg und benötigten 5 Tage zur Heilung, 1 Patient erhielt 3 mg und benötigte 6 Tage zur Heilung, 5 Patienten erhielten 4 mg und benötigten 3 Tage zur Heilung und 2 Patienten erhielten 4 mg und benötigten 5 Tage zur Heilung.

a) Berechnen Sie die Regressionsgerade der Heilungstage in Abhängigkeit von der Dosis.
b) Berechnen Sie den Regressionskoeffizienten der Heilungstage in Abhängigkeit von der Dosis und interpretieren Sie ihn.
c) Berechnen Sie den Korrelationskoeffizienten und interpretieren Sie ihn.
d) Bestimmen Sie die erwartete Zeit, die für die Heilung mit einer Dosis von 5 mg benötigt wird. Ist diese Vorhersage zuverlässig? Begründen Sie die Antwort.
e) Welche Dosis muss angewendet werden, um in 4 Tagen zu heilen? Ist diese Vorhersage zuverlässig? Begründen Sie die Antwort.

Für die Lösung siehe Abschnitt 2.21

1.3.7 Aufgabe 1.3.7 Gewicht und Körpergröße

Im Datensatz `heigths.weights.students` von rk.Teaching[6] sind Gewicht und Körpergröße von 100 Probanden enthalten.

a) Laden Sie den Datensatz `heigths.weights.students` in Ihre R-Session.
b) Führen Sie eine lineare Regression `Gewicht erklärt durch Größe` durch.
c) Erstellen Sie eine Punktwolke inklusive Regressionsgeraden jeweils für Männer und Frauen getrennt.
d) Berechnen Sie die Bestimmtheitskoeffizienten (R^2) für beide Modelle. Welches Modell erklärt besser die Beziehung zwischen Gewicht und Größe, das der Männer oder das der Frauen? Begründen Sie die Antwort.
e) Was ist das zu erwartende Gewicht für einen Mann mit 170cm Körpergröße? Und für eine Frau der selben Größe?

Für die Lösung siehe Abschnitt 2.22

1.3.8 Aufgabe 1.3.8 Neugeborene

Der Datensatz `neonates` von rk.Teaching[7] enthält Informationen über eine Stichprobe von 320 Neugeborenen, die im Laufe eines Jahres nach normaler Schwangerschaftsdauer geboren wurden.

a) Erstellen Sie eine Kreuztabelle vom APGAR-Wert nach 1 Minute und dem Rauchverhalten der Mütter während der Schwangerschaft. Welche Schlüsse lassen sich ziehen?
b) Erstellen Sie eine Kreuztabelle vom APGAR-Wert nach 1 Minute und der Alterskategorie der Mütter. Welche Schlüsse lassen sich ziehen?
c) Führen Sie eine lineare Regression für `Geburtsgewicht erklärt durch Anzahl täglich gerauchter Zigaretten` durch. Gibt es einen starken linearen Zusammenhang?
d) Plotten Sie Ihre Regression. Passt die Regressionsgerade gut zur Punktwolke?
e) Wiederholen Sie die Regression, aber nutzen Sie dieses Mal nur Daten von Raucherinnen. Ist dieses Modell besser oder schlechter als das vorherige? Wieviel Gewicht verliert ein Neugeborenes nach diesem Modell pro täglich gerauchter Zigarette?
f) Welches Geburtsgewicht sagt dieses Modell für ein Neugeborenes vorher, dessen Mutter 5 Zigaretten täglich während der Schwangerschaft geraucht hat? Wieviel für eine Mutter, die 30 Zigaretten täglich raucht. Wie zuverlässich sind diese Ergebnisse?
g) Ändert sich der lineare Zusammenhang, wenn die Daten nach Altersgruppen getrennt untersucht werden?

Für die Lösung siehe Abschnitt 2.23

[6] https://github.com/rkward-community/rk.Teaching, auch verfügbar unter https://www.produnis.de/R/data/heights.weights.students.RData
[7] https://github.com/rkward-community/rk.Teaching, auch verfügbar unter https://www.produnis.de/R/data/neonates.RData

1.4 Nicht-lineare Regression

1.4.1 Aufgabe 1.4.1 Bakterien

Die Anzahl an Bakterien in einer Kultur vermehrt sich wie folgt:

Stunden	1	2	3	4	5	6	7	8	9
Bakterien	25	28	47	65	86	121	190	290	362

a) Erstellen Sie ein Datenframe mit den Variablen `Stunden` und `Bakterien`.
b) Erzeugen Sie ein Scatterplot. Welche Regression würden Sie auf Grundlage des Plots vorschlagen?
c) Berechnen Sie die quadratischen und exponentiellen Modelle für die Bakterienvermehrung über die Zeit.
d) Plotten Sie das bessere Modell in die Punktwolke.
e) Wie viele Bakterien werden nach dem besten Modell 3 Stunden nach Anlegen der Kultur vorhanden sein? Und nach 10 Stunden? Sind diese Vorhersagen zuverlässig?
f) Machen Sie eine möglichst zuverlässige Vorhersage über die Zeit, die benötigt wird, um 100 Bakterien in der Kultur zu haben.

Für die Lösung siehe Abschnitt 2.24

1.4.2 Aufgabe 1.4.2 Diät

Der Datensatz `diet` von rk.Teaching[8] enthält Informationen über eine Diätenuntersuchung. Für jede Person wurde die Anzahl der Diättage, der Gewichtsverlust und die regelmäßige körperliche Betätigung gemessen.

a) Laden Sie den Datensatz `diet` in Ihre R-Session.
b) Erstellen Sie eine Punktwolke. Welche Art von Modell erklärt auf Grundlage der Punktwolke den Gewichtsverlust pro Diättag besser?
c) Berechnen Sie das Regressionsmodell, welches den Gewichtsverlust mit der Anzahl an Diättagen am besten (im Vergleich zu anderen) erklären kann. Wird das Modell zuverlässige Vorhersagen machen?
d) Plotten Sie Ihr Modell.
e) Berechnen Sie das Regressionsmodell, das den Gewichtsverlust mit den Tagen der Diät für die Gruppe der Personen, die sich nicht regelmäßig körperlich betätigen, am besten erklären kann.
f) Wiederholen Sie die Analyse für die Gruppe, die sich regelmäßig körperlich betätigt.
g) Benutzen Sie die erstellen Modelle, um den Gewichtsverlust nach 30 und nach 100 Tagen Diät für Personen, die sich körperlich betätigen, und für solche, die dies nicht tun, vorherzusagen. Sind diese Vorhersagen zuverlässig?

[8] https://github.com/rkward-community/rk.Teaching, auch verfügbar unter https://www.produnis.de/R/data/diet.RData

Für die Lösung siehe Abschnitt 2.25

1.4.3 Aufgabe 1.4.3 Blutkonzentration

Die Konzentration eines Arzneimittels im Blut in mg/dl hängt von der Zeit ab, wie aus den folgenden Daten hervorgeht.

Stunden	2	3	4	5	6	7	8
Konzentration	25	36	48	64	86	114	168

a) Benutzen Sie ein exponentielles Modell, um die Konzentration nach 10 Stunden vorherzusagen. Ist die Vorhersage zuverlässig?
b) Benutzen Sie ein logarithmisches Modell um zu bestimmen, nach wie vielen Stunden eine Konzentration von 100 mg/dl erreicht sein wird.

Für die Lösung siehe Abschnitt 2.26

1.5 Wahrscheinlichkeiten

1.5.1 Aufgabe 1.5.1 Glücksspiel

Lassen Sie in R ...

a) eine beliebige Poker-Spielkarte[9] ziehen.
b) 2 Münzen werfen.
c) 2 Würfeln werfen.

Für die Lösung siehe Abschnitt 2.27

1.5.2 Aufgabe 1.5.2 Münzwürfe

Wiederholen Sie die Zufallsexperimente und lassen Sie R 10 mal, 100 mal 1.000 mal und 1.000.000 mal zwei Münzen werfen.

a) Erstellen Sie je eine relative Häufigkeitstabelle der Ergebnisse. Wie sind die Tabellen zu

[9] Den Datensatz für ein Pokerkartenspiel erhalten Sie unter https://www.produnis.de/R/data/cards.RData

bewerten?
b) Welche theoretischen Wahrscheinlichkeiten haben die möglichen Wurfergebnisse? Stimmen diese mit den beobachteten Ergebnissen überein?

Für die Lösung siehe Abschnitt 2.28

1.5.3 Aufgabe 1.5.3 Medizinschrank

In einem Medizinschrank befinden sich drei Boxen mit Medikament A, zwei Boxen mit Medikament B und eine Box mit Medikament C.

a) Ziehen Sie zufällig 3 Boxen, ohne zurücklegen.
b) Ziehen Sie zufällig 3 Boxen, diesmal mit zurücklegen.

Für die Lösung siehe Abschnitt 2.29

1.5.4 Aufgabe 1.5.4 Kinderkrankheiten

Eine epidemiologische Untersuchung wurde durchgeführt, um die Lebenszeitprävalenz von drei häufigen Kinderkrankheiten zu ermitteln: Windpocken, Masern und Röteln. Die beobachteten Häufigkeiten sind in der nachstehenden Tabelle aufgeführt.

Windpocken	Masern	Röteln	Häufigkeit
No	No	No	2654
No	No	Yes	1436
No	Yes	No	1682
No	Yes	Yes	668
Yes	No	No	1747
Yes	No	Yes	476
Yes	Yes	No	876
Yes	Yes	Yes	265

a) Erstellen Sie ein Datenframe mit den Variablen `Windpocken`, `Masern`, `Röteln` und `Häufigkeit` und übertragen Sie die Daten.
b) Erstellen Sie den Wahrscheinlichkeitsraum der Lebenszeitprävalenz.
c) Wie groß ist die Wahrscheinlichkeit, dass eine zufällig gezogene Person Windpocken hatte?
d) Wie groß ist die Wahrscheinlichkeit, dass eine zufällig gezogene Person Windpocken oder Masern hatte?
e) Wie groß ist die Wahrscheinlichkeit, dass eine zufällig gezogene Person Masern und Röteln hatte?
f) Wie groß ist die Wahrscheinlichkeit, dass eine zufällig gezogene Person, die bereits an Masern erkrankte, nun an Windpocken erkrankt?
g) Wie groß ist die Wahrscheinlichkeit, dass eine zufällig gezogene Person, die keine Masern und keine Röteln hatte, an Windpocken erkrankt?

Für die Lösung siehe Abschnitt 2.30

1.5.5 Aufgabe 1.5.5 Schwangerschaftstest

Ein Schwangerschaftstest, der von vielen Frauen angewendet wurde, erzielte folgende Ergebnisse.

Schwanger	Test	Häufigkeit
Nein	-	3876
Nein	+	47
Ja	-	12
Ja	+	131

a) Erstellen Sie ein Datenframe mit den Variablen `Schwanger`, `Testergebnis` und `Häufigkeit`.
b) Erstellen Sie den Wahrscheinlichkeitsraum.
c) Berechnen Sie die Prävalenz der Schwangerschaften.
d) Wie groß ist die Wahrscheinlichkeit, ein positives Testergebnis zu ziehen?
e) Bestimmen Sie die Sensitivität des Tests
f) Bestimmen Sie die Spezifität des Tests
g) Bestimmen Sie den positiv prädiktiven Wert des Tests
h) Bestimmen Sie den negativ prädiktiven Wert des Tests

Für die Lösung siehe Abschnitt 2.31

1.5.6 Aufgabe 1.5.6 Glückspielwahrscheinlichkeiten

Erstelle den Ereignisraum des Zufallsexperiments, das aus dem Werfen einer Münze, dem Werfen eines Würfels und dem Ziehen einer Karte aus einem spanischen Kartenspiel besteht.

Für die Lösung siehe Abschnitt 2.32

1.5.7 Aufgabe 1.5.7 Grippeimpfung

Die Wirksamkeit eines Grippeimpfstoffs wurde an 1.000 Probanden erprobt.

Impfung	Grippe	Häufigkeit
Nein	Nein	418
Nein	Ja	312
Ja	Nein	233
Ja	Ja	37

a) Erzeugen Sie den Wahrscheinlichkeitsraum
b) Wie groß ist die Wahrscheinlichkeit, dass eine zufällig ausgewählte Person geimpft ist?
c) Wie hoch ist die Prävalenz der Grippe?
d) Wie groß ist die Wahrscheinlichkeit, dass geimpfte Personen an Grippe erkranken? Ist die Impfung effektiv?

Für die Lösung siehe Abschnitt 2.33

1.5.8 Aufgabe 1.5.8 Ebola

Um die Wirksamkeit eines Diagnosetests zur Feststellung von Ebola in einem zentralafrikanischen Land zu ermitteln, wurde der Test an vielen Personen durchgeführt. Das Ergebnis des Tests war positiv bei 147 Personen mit Ebola, aber aber auch bei 28 Personen ohne Ebola. Negativ war das Ergebnis des Tests bei 97465 Personen ohne Ebola, aber auch bei 65 Personen mit Ebola.

a) Erzeugen Sie den Wahrscheinlichkeitsraum des Tests.
b) Berechnen Sie die Prävalenz von Ebola in der Bevölkerung.
c) Wie hoch ist die Wahrscheinlichkeit, ein negatives Testergebnis zu erhalten?
d) Berechnen Sie die Sensitivität und Spezifität des Tests.
e) Kann der Test besser Erkrankte erkennen, oder Gesunde?
f) Wenn eine Person einen positiven Test erhält, wie hoch ist dann die Wahrscheinlichkeit, dass er tatsächlich krank ist?
g) Wenn eine Person einen negativen Test erhält, wie hoch ist dann die Wahrscheinlichkeit,

dass er tatsächlich gesund ist?

Für die Lösung siehe Abschnitt 2.34

1.6 Diskrete Wahrscheinlichkeitsverteilungen

1.6.1 Aufgabe 1.6.1 Münzwurf

Wir haben 10 mal eine Münze geworfen, wobei das Ergebnis der Binomialverteilung B(10;0.5) folgt. Die Variable `X` misst, wie häufig dabei "`Kopf`" geworfen wurde.

a) Berechnen Sie die Wahrscheinlichkeitsverteilung von `X`
b) Plotten Sie die Wahrscheinlichkeitsfunktion von `X`
c) Plotten Sie die Verteilungsfunktion.
d) Berechnen Sie die Wahrscheinlichkeit, 7 mal `Kopf` zu werfen.
e) Berchnen Sie die Wahrscheinlichkeit, weniger als als 4 mal `Kopf` zu werfen.
f) Berchnen Sie die Wahrscheinlichkeit, mehr als als 5 mal `Kopf` zu werfen.
g) Berechnen Sie die Wahrscheinlichkeit, 2 bis 8 mal `Kopf` zu werfen.

Für die Lösung siehe Abschnitt 2.35

1.6.2 Aufgabe 1.6.2 Geburten pro Tag

Die Anzahl an täglichen Geburten `X` in unserer Stadt folgt einer Poissonverteilung mit durchschnittlich 6 Geburten am Tag.

a) Plotten Sie die Wahrscheinlichkeitsfunktion von `X`
b) Plotten Sie die Verteilungsfunktion von `X`
c) Wie groß ist die Wahrscheinlicheit, dass an einem zufälligen Tag (nur) 1 Geburt stattfindet?
d) Wie groß ist die Wahrscheinlicheit, dass an einem zufälligen Tag weniger als 6 Geburten stattfinden?
e) Wie groß ist die Wahrscheinlicheit, dass an einem zufälligen Tag 4 oder mehr Geburten stattfinden?
f) Wie groß ist die Wahrscheinlicheit, dass an einem zufälligen Tag 4 bis 8 Geburten stattfinden?
g) Wie groß ist die Wahrscheinlicheit, dass in einer Woche zwischen 30 und 40 Geburten stattfinden?

Für die Lösung siehe Abschnitt 2.36

1.6.3 Aufgabe 1.6.3 Gesetz der seltenen Ereignisse

Kommen wir nochmal auf das Münzwurfbeispiel aus Abschnitt 1.6.1 zurück.
Das Gesetz der seltenen Ereignisse besagt, dass das Binomial-Verteilungsmodell $B(n, p)$ zum Poisson-Wahrscheinlichkeitsverteilungsmodell $P(np)$ tendiert, wenn n gegen ∞ und p gegen 0 tendiert. Insbesondere ist das Poisson-Modell eine gute Annäherung an das Binomialmodell für $n \geq 30$ und $p \leq 0,1$.
Zur Überprüfung dieses Gesetz,

a) berechnen Sie die Wahrscheinlichkeitsverteilung des binomialen Modells $B(30, 0.1)$.
b) berechnen Sie die Wahrscheinlichkeitsverteilung des Poissonmodells $P(3)$ und vergleichen Sie es mit dem binomialen Modell $B(30, 0.1)$.
c) berechnen Sie die Wahrscheinlichkeitsverteilung des binomialen Modells $B(100, 0.3)$ und vergleichen Sie es es mit dem Modell $P(3)$. Sind diese Modelle ähnlicher als die vorherigen?
d) Plotten Sie die Wahrscheinlichkeitsfunktionen der vorherigen Modelle. Erhöhen Sie die Anzahl der Wiederholungen und verringern Sie die Erfolgswahrscheinlichkeit im Binomialmodell und beobachten Sie, wie sich die Wahrscheinlichkeiten des Binomialmodells und des Poissonmodells annähern.

Für die Lösung siehe Abschnitt 2.37

1.6.4 Aufgabe 1.6.4 Münzwürfe (II)

Wie groß ist die Wahrscheinlichkeit, beim Werfen von 100 Münzen zwischen 40 und 60 Mal `Kopf` zu erhalten (beide Werte eingeschlossen)?

Für die Lösung siehe Abschnitt 2.38

1.6.5 Aufgabe 1.6.5 Behandlungserfolg

Die Wahrscheinlichkeit, dass eine Behandlung Erfolg hat, liegt bei 85%. Wenn wir an 6 Personen die Behandlung durchführen,

a) wie groß ist die Wahrscheinlichkeit, dass die Hälfte der Patienten geheilt wird?
b) wie groß ist die Wahrscheinlichkeit, dass mindestens 4 Patienten geheilt werden?
c) plotten Sie die Wahrscheinlichkeitsfunktion für die Anzahl geheilter Patienten.

Für die Lösung siehe Abschnitt 2.39

1.6.6 Aufgabe 1.6.6 Impfreaktion

Die Wahrscheinlichkeit einer starken Impfreaktion beträgt $0,001$. Wenn 2.000 Personen geimpft werden, wie hoch ist die Wahrscheinlichkeit für starke Reaktionen?

Für die Lösung siehe Abschnitt 2.40

1.6.7 Aufgabe 1.6.7 Telefonanrufe

Die durchschnittliche Anzahl an Telefonanrufen in unserer Telefonzentrale beträgt 120 Anrufe pro Minute.

a) Wie hoch ist die Wahrscheinlichkeit, dass weniger als 4 Anrufe in 2 Sekunden eintreffen?
b) Wie hoch ist die Wahrscheinlichkeit, dass mindestens 3 Anrufe in 3 Sekunden eintreffen?

Für die Lösung siehe Abschnitt 2.41

1.7 Kontinuierliche Wahrscheinlichkeitsverteilungen

1.7.1 Aufgabe 1.7.1 Bushaltestelle

Nehmen wir an, dass ein Bus alle 15 Minuten an einer Haltestelle vorbeifährt und dass eine Person zu jedem Zeitpunkt mit der gleichen Wahrscheinlichkeit eintreffen kann. Dann folgt die Variable, die die Wartezeit auf den Bus misst, einer gleichmäßigen Wahrscheinlichkeitsverteilung $U(0, 15)$, da jede Wartezeit zwischen 0 und 15 Minuten die gleiche Wahrscheinlichkeit hat.

a) Plotten Sie die Dichtefunktion der Wartezeit.
b) Plotten Sie die Verteilungsfunktion der Wartezeit.
c) Berechnen Sie die Wahrscheinlichkeit, weniger als 5 Minuten auf den Bus zu warten.
d) Berechnen Sie die Wahrscheinlichkeit, länger als 12 Minuten auf den Bus zu warten.
e) Berechnen Sie die Wahrscheinlichkeit, zwischen 5 und 10 Minuten auf den Bus zu warten.
f) Bei welcher Zeit zwischen 0 und 15 Minuten muss die Hälft der Personen kürzer auf den Bus warten als die angegebene Zeit?
g) Bei welcher Zeit zwischen 0 und 15 Minuten müssen 10% der Personen länger auf den Bus warten als die angegebene Zeit?

Für die Lösung siehe Abschnitt 2.42

1.7.2 Aufgabe 1.7.2 Standardnormalverteilung

Eine Variable folgt in ihren Ausprägungen der Standardnormalverteilung ($Z \sim N(0, 1)$)

a) Plotten Sie die Dichtefunktion von Z.
b) Wie beeinflussen Mittelwert und Standardabweichung die Form der Gausschen Glockenkurve?
c) Plotten Sie die Verteilungsfunktion von Z.
d) Berechnen Sie die Wahrscheinlichkeit $P(Z < -1)$.
e) Berechnen Sie die Wahrscheinlichkeit $P(Z > 1)$
f) Berechnen Sie die Wahrscheinlichkeit, dass Z zwischen dem Mittelwert minus der Standardabweichung und dem Mittelwert plus der Standardabweichung liegt, d. h. $P(-1 \leq Z \leq 1)$.
g) Berechnen Sie die Wahrscheinlichkeit, dass Z zwischen dem Mittelwert minus zwei Standardabweichungen und dem Mittelwert plus zwei Standardabweichungen liegt, d. h. $P(-2 \leq Z \leq 2)$.
h) Berechnen Sie die Wahrscheinlichkeit, dass Z zwischen dem Mittelwert minus drei Standardabweichungen und dem Mittelwert plus drei Standardabweichungen liegt, d. h. $P(-3 \leq Z \leq 3)$.
i) Berechnen Sie die Quartile.
j) Bei welchem Z-Wert liegen 95% der Fläche unterhalb des Wertes?
k) Bei welchem Z-Wert liegen 2,5% der Fläche oberhalb des Wertes?

Für die Lösung siehe Abschnitt 2.43

1.7.3 Aufgabe 1.7.3 Chiquadratverteilungen

Wenn $X_1, \ldots, X_n$ unabhängige standardnormalverteilte Werte sind, dann folgt die Variable $X = X_1^2 + \cdots + X_n^2$ einer Chiquadratverteilung mit n Freiheitsgraden ($\chi^2(n)$). Nehmen wir nun an, X würde der Chiquadratverteilung mit 6 Freiheitsgraden folgen ($\chi^2(6)$).

a) Plotten Sie die Dichtefunktion dieser Verteilung
b) Wie groß ist die Wahrscheinlichkeit für $P(X < 6)$?
c) Berechnen Sie das fünfte Perzentil der Verteilung.
d) Bei welchem Wert liegen 10% der Fläche oberhalb des Wertes?

Für die Lösung siehe Abschnitt 2.44

1.7.4 Aufgabe 1.7.4 t-Verteilung

Wenn Y einer Chiquadratverteilung mit n Freiheitsgraden folgt ($\chi^2(n)$) und Z der Standardnormalverteilung ($N(0,1)$), dann folgt die Variable $X = \frac{Z}{\sqrt{Y/n}}$ einer Student-t-Verteilung mit 8 Freiheitsgraden ($T(8)$).

a) Plotten Sie die Dichtefunktion von X und vergleichen Sie diese mit der Dichtefunktion der Standardnormalverteilung.
b) Berechnen Sie das 8te Perzentil von X.
c) Bei welchem Wert von X liegen 5% aller Fälle oberhalb dieses Wertes?

Für die Lösung siehe Abschnitt 2.45

1.7.5 Aufgabe 1.7.5 Fishers F-Verteilung

Wenn Y_1 und Y_2 zwei unabhängige Variablen aus den Chiquadratverteilungen mit n und m Freiheitsgraden stammen, dann folgt die Variable $X = \frac{Y_1/n}{Y_2/m}$ einer Fisher-F-Verteilung mit n und m Freiheitsgraden ($F(n,m)$). Nehmen wir an, X folge einer Fisher-F-Verteilung mit 10 und 20 Freiheitsgeraden ($F(10,20)$).

a) Plotten Sie die Dichtefunktion von X.
b) Berechnen Sie Wahrscheinlichkeit $P(X > 1)$.
c) Berechnen Sie den Interquartilsabstand.

Für die Lösung siehe Abschnitt 2.46

1.7.6 Aufgabe 1.7.6 Blutzuckerspiegel

Es ist bekannt, dass der Glukosespiegel im Blut von Diabetikern einem Normalverteilungsmodell mit einem Mittelwert von 106 mg/100 ml und einer Standardabweichung von 8 mg/100 ml folgt.

a) Berechnen Sie die Wahrscheinlichkeit, dass ein zufällig ausgewählter Diabetiker einen Glukosespiegel von weniger als 120 mg/100 ml hat.
b) Wie viel Prozent der Personen haben einen Glukosespiegel zwischen 90 und 120 mg/100 ml?
c) Berechnen und interpretieren Sie das erste Quartil des Glukosespiegels.

Für die Lösung siehe Abschnitt 2.47

1.7.7 Aufgabe 1.7.7 Cholesterinspiegel bei Männern

Es ist bekannt, dass der Cholesterinspiegel bei Männern im Alter von 30 Jahren einer Normalverteilung folgt mit Mittelwert 220 mg/dl und einer Standardabweichung von 30 mg/dl. In einer bestimmten Population gibt es 20.000 Männer im Alter von 30 Jahren.

a) Wie viele von ihnen haben einen Cholesterinspiegel zwischen 210 und 240 mg/dl?
b) Wenn ein Cholesterinspiegel von mehr als 250 mg/dl eine Thrombose auslösen kann, wie viele von ihnen sind thrombosegefährdet?
c) Welcher Cholesterinwert wird von mindestens 20% der Männer erreicht?

Für die Lösung siehe Abschnitt 2.48

1.8 Konfidenzintervalle (eine Stichprobe)

1.8.1 Aufgabe 1.8.1 Wirkstoffkonzentration

Die Wirkstoffkonzentration einer Zufallsstichprobe von 10 Arzneimittelbehältern aus einer Charge beträgt (in mg/mm^3)

17.6 19.2 21.3 15.1 17.6 18.9 16.2 18.3 19.0 16.4

a) Übertragen Sie die Daten in ein Datenframe mit der Variable `Konzentration`.
b) Berechnen Sie das Konfidenzintervall für die mittlere Konzentration bei einem Konfidenzniveau von 95% (Signifikanzlevel $\alpha = 0,05$).
c) Berechnen Sie das Konfidenzintervall für die mittlere Konzentration bei einem Konfidenzniveau von 99% (Signifikanzlevel $\alpha = 0,01$).
d) Wenn wir die Genauigkeit des Intervalls als den Kehrwert seiner Breite definieren, wie ändert sich die Genauigkeit eines Intervalls, wenn wir das Konfidenzniveau erhöhen? Warum?
e) Welche Stichprobengröße wird benötigt, um den mittleren Konzentrationswert mit einem Fehler von $\pm 0.5 mg/mm^3$ und einem Konfidenzniveau von 95% Sicherheit zu bestimmen?
f) Wenn die Konzentration des Wirkstoffs mindestens 16 mg/mm^3 betragen muss, um wirksam zu sein, ist dann unsere Medikamentencharge wirksam?

Für die Lösung siehe Abschnitt 2.49

1.8.2 Aufgabe 1.8.2 Milchfett

Ein Molkereibetrieb erhält Milch von zwei Bauernhöfen X und Y. Um die Qualität der Milch zu analysieren, wird das Milchfett für zwei Milchproben, eine von jedem Betrieb, gemessen. Die Ergebnisse sind in der nachstehenden Tabelle aufgeführt.

X		Y	
0.34	0.34	0.28	0.29
0.32	0.35	0.30	0.32
0.33	0.33	0.32	0.31
0.32	0.32	0.29	0.29
0.33	0.30	0.31	0.32
0.31	0.32	0.29	0.31
		0.33	0.32
		0.32	0.33

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `Hof1` und `Hof2`.
b) Berechnen Sie das 95%-Konfidenzintervall für den durchschnittlichen Fettgehalt.
c) Berechnen Sie das 95%-Konfidenzintervall für den durchschnittlichen Fettgehalt, getrennt nach Höfen.
d) Plotten Sie das 95%-Konfidenzintervall für den durchschnittlichen Fettgehalt, getrennt nach Höfen.
e) Lässt sich aus den Konfidenzintervallen ein signifikanter Untschied zwischen den Höfen feststellen?

Für die Lösung siehe Abschnitt 2.50

1.8.3 Aufgabe 1.8.3 Bibliotheksnutzung

In einer von einer Universität durchgeführten Umfrage über die Nutzung der Bibliothek wurde eine Stichprobe von 34 Studierenden gefragt, ob sie mindestens einmal pro Woche in die Bibliothek gehen.

nein ja nein nein nein ja nein ja ja ja ja nein ja nein ja nein nein nein ja ja ja nein
nein ja nein nein ja ja nein nein ja nein ja nein

a) Übertragen Sie die Daten in ein Datenframe mit der Variable `Antwort`.
b) Berechnen Sie das Konfidenzintervall für den Anteil an Studierenden, welche die Bibliothek wöchentlich nutzen mit einem Signifikanzlevel von $\alpha = 0,01$.
c) Wie präzise ist das Intervall?
d) Welcher Stichprobenumfang ist erforderlich, um eine Schätzung des Anteils der Studierenden zu erhalten, die die Bibliothek mindestens einmal pro Woche nutzen, mit einem Fehler von $\pm 1\%$ und einem Konfidenzniveau von 95%?

Für die Lösung siehe Abschnitt 2.51

1.8.4 Aufgabe 1.8.4 Atemwegsprobleme und Impfung

Das Gesundheitsministerium möchte ein Konfidenzintervall für den Anteil der Personen über 65 Jahre mit Atemwegsproblemen berechnen, die geimpft worden sind. In einer Zufallsstichprobe von 200 Personen über 65 mit Atemwegsproblemen wurden 154 geimpft.

a) Berechnen Sie das 95%-Konfidenzintervall für den Anteil an geimpften Probanden in der Grundgesamtheit.
b) Wenn das Gesundheitsministerium das Ziel verfolgt, dass mindestens 70% der Menschen über 65 mit Atemwegserkrankungen geimpft sind, können wir dann sagen, dass das Ministerium das Ziel erreicht hat?

Für die Lösung siehe Abschnitt 2.52

1.8.5 Aufgabe 1.8.5 Cholesterin

Der Cholesterinspiegel (in mg/dl) in einer Zufallsstichprobe mit 8 Probanden beträgt

196 212 188 206 203 210 201 198

a) Berechnen Sie die Konfidenzintervalle für den Mittelwert mit den Signifikanzniveaus 0.1, 0.05 und 0.01.

B) Kann man schließen, dass der Mittelwert des Cholesterinspiegels der Bevölkerung unter 210 mg/dl liegt?

Für die Lösung siehe Abschnitt 2.53

1.8.6 Aufgabe 1.8.6 Neurologisches Syndrom

Zur Behandlung eines neurologischen Syndroms gibt es zwei Therapien, A und B. In einer Studie wurde eine Stichprobe von 60 Personen gezogen. Bei 25 von ihnen wurde Therapie A angewandt, bei den anderen 35 Therapie B. Insgesmant 18 der mit A behandelten Personen wurden geheilt, während 21 der mit B behandelten Personen geheilt wurden.

a) Berechnen Sie für jede Therapie das 95% Konfidenzintervall für den Anteil an Personen,

die geheilt wurden.
b) Welches Intervall ist präziser?

Für die Lösung siehe Abschnitt 2.54

1.8.7 Aufgabe 1.8.7 Neugeborene

Der Datensatz `neonates` von rk.Teaching[10] enthält Informationen über eine Stichprobe von 320 Neugeborenen, die im Laufe eines Jahres nach normaler Schwangerschaftsdauer geboren wurden.

a) Berechnen Sie das 99% Konfidenzintervall für den Mittelwert des Gewichts der Neugeborenen.
b) Berechnen Sie die Konfidenzintervalle für den APGAR-Score nach 1 Minute und für den APGAR-Score nach 5 Minuten und vergleiche sie beide Intervalle. Gibt es auf Grundlage der Konfidenzintervalle einen signifikanten Unterschied zwischen den Mittelwerten der beiden Scores?
c) Berechnen Sie die Konfidenzintervalle für den Prozentsatz der Neugeborenen mit einem Gewicht von $\leq 2,5$ kg für Raucher- und Nichtrauchermütter und vergleichen Sie die Intervalle.

Für die Lösung siehe Abschnitt 2.55

1.9 Konfidenzintervalle (zwei Stichproben)

1.9.1 Aufgabe 1.9.1 Medikamentenwerbung

Um festzustellen, ob eine Werbekampagne den Absatz eines Arzneimittels erhöht hat, wurde eine Stichprobe von 8 Apotheken aus einer Stadt gezogen. In jeder Apotheke wurden die monatlichen Verkäufe des Arzneimittels vor und nach der Kampagne in der folgenden Tabelle erfasst.

Vorher	147	163	121	205	132	190	176	147
Nachher	150	171	132	208	141	184	182	145

[10] https://github.com/rkward-community/rk.Teaching, auch verfügbar unter https://www.produnis.de/R/data/neonates.RData

a) Erstellen Sie ein Datenframe mit den Variablen `vorher` und `nachher` und übertragen Sie die Daten.
b) Berechnen Sie den Mittelwert der monatlichen Umsätze vor und nach der Kampagne. Sind die Mittelwerte unterschiedlich? Hat die Kampagne den Absatz des Arzneimittels erhöht?
c) Berechnen Sie die Konfidenzintervalle für den durchschnittlichen Unterschied mit $\alpha = 0,05$ und $\alpha = 0,01$.
d) Können wir dieselbe Schlussfolgerung ziehen, wenn wir die Verkäufe nach der Kampagne der beiden letzten Apotheken ändern und 190 statt 182 und 165 statt 145 angeben? Was passiert mit den Konfidenzintervallen?

Für die Lösung siehe Abschnitt 2.56

1.9.2 Aufgabe 1.9.2 Milchfett

Ein Molkereibetrieb erhält Milch von zwei Bauernhöfen X und Y. Um die Qualität der Milch zu analysieren, wird das Milchfett für zwei Milchproben, eine von jedem Betrieb, gemessen. Die Ergebnisse sind in der nachstehenden Tabelle aufgeführt.

X		Y	
0.34	0.34	0.28	0.29
0.32	0.35	0.30	0.32
0.33	0.33	0.32	0.31
0.32	0.32	0.29	0.29
0.33	0.30	0.31	0.32
0.31	0.32	0.29	0.31
		0.33	0.32
		0.32	0.33

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `Hof1` und `Hof2`.
b) Berechnen Sie das 95%-Konfidenzintervall für den durchschnittlichen Fettunterschied in der Milch von `Hof1` und `Hof2`.
c) Kann man daraus schließen, dass der Unterschied zwischen den Milchfettmittelwerten der Betriebe signifikant ist? Welcher Betrieb hat Milch mit mehr Fett? Wie viel mehr Fett hat die Milch von `Hof1` als die Milch von `Hof2`?

Für die Lösung siehe Abschnitt 2.57

1.9.3 Aufgabe 1.9.3 Bibliotheksnutzung nach Geschlecht

In einer von einer Universität durchgeführten Umfrage über die Nutzung der Bibliothek wurde eine Stichprobe von 34 Studierenden gefragt, ob sie mindestens einmal pro Woche in die Bibliothek gehen.

Antwort	nein	ja	nein	nein	nein	ja	nein	ja	ja	ja	ja	nein
Geschlecht	m	w	w	m	m	m	w	w	w	w	m	m

Antwort	nein	ja	nein	nein	nein	ja	ja	ja	nein	nein	ja	nein
Geschlecht	m	w	m	m	w	m	w	w	w	m	w	m

Antwort	ja	ja	nein	nein	ja	nein	ja	nein	ja	nein
Geschlecht	w	w	m	m	w	w	w	m	w	m

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `Antwort` und `Geschlecht`.
b) Berechnen Sie das Konfidenzintervall für den Unterschied zwischen den Anteilen der Frauen und Männern, die die Bibliothek mindestens einmal pro Woche nutzen.

Für die Lösung siehe Abschnitt 2.58

1.9.4 Aufgabe 1.9.4 Prüfungen vormittags und nachmittags

In einem Kurs gibt es zwei Gruppen von Studierenden, eine am Vormittag und die andere am Nachmittag. In der Vormittagsgruppe haben 55 von 80 Studierenden bestanden, während in der Nachmittagsgruppe 32 von 90 Studierenden bestanden haben.

a) Gibt es signifikante Unterschiede zwischen den Prozentsätzen der Studierenden, die am Vormittag und am Nachmittag bestanden haben? Kann man daraus schließen, dass der Stundenplan die Ursache für diese Unterschiede ist?

Für die Lösung siehe Abschnitt 2.59

1.9.5 Aufgabe 1.9.5 Cholesterin und Sport

In einer Studie zur Ermittlung des Zusammenhangs zwischen körperlicher Betätigung und dem Cholesterinspiegel im Blut wurde eine wurde eine Stichprobe von 11 Personen gezogen. Der Cholesterinspiegel der Teilnehmer (in mg/dl) vor und nach der Teilnahme an einem Programm mit körperlichen Übungen ist unten dargestellt.

vorher	182	232	191	200	148	249	276	213	241	280	262
nachher	198	210	194	220	138	220	219	161	210	213	226

a) Berechnen Sie das 95%-Konfidenzintervall für den durchschnittlichen Unterschied der Cholesterinwerte vor und nach den körperlichen Übungen
b) Berechnen Sie das 99%-Konfidenzintervall für den durchschnittlichen Unterschied der Cholesterinwerte vor und nach den körperlichen Übungen
c) Auf Grundlage der zuvor berechneten Intervalle, welchen Schluss bezüglich des Einflusses von körperlichen Aktivitäten auf den Cholesterinspiegel können Sie ziehen?

Für die Lösung siehe Abschnitt 2.60

1.9.6 Aufgabe 1.9.6 Patientenzufriedenheit

Insgesamt 500 Patienten aus zwei Krankenhäusern wurden zu ihrer Zufriedenheitbefragt. In Krankenhaus 1 wurden 200 Patienten befragt, von denen 140 zufrieden waren. In Krankenhaus 2 wurden 300 Patienten befragt, von denen 180 zufrieden waren.

a) Berechnen Sie das 95%-Konfidenzintervall für den Anteilsunterschied an zufriedenen Patienten in beiden Häusern.
b) Wenn $\alpha = 0,01$ ist, können dann Rückschlüsse gezogen werden, ob der Unterschied der Anteile zufriedener Patienten signifikant ist?

Für die Lösung siehe Abschnitt 2.61

1.9.7 Aufgabe 1.9.7 Neugeborene

Der Datensatz `neonates` von rk.Teaching[11] enthält Informationen über eine Stichprobe von 320 Neugeborenen, die im Laufe eines Jahres nach normaler Schwangerschaftsdauer geboren wurden.

a) Berechnen Sie das 95%-Konfidenzintervall für den durchscnnittlichen Unterschied des Geburtsgewichts zwischen Kindern von Raucherinnen und Nichtraucherinnen. Wie groß ist der durchschnittliche Gewichtsunterschied?

b) Berücksichtigen Sie nur die Daten der Mütter, die *während* der Schwangerschaft nicht geraucht haben. Berechnen Sie das 95%-Konfidenzintervall für den durchscnnittlichen Unterschied des Geburtsgewichts zwischen Kindern von Müttern, die *vor* der Schwangerschaft geraucht haben, und den Nichtraucherinnen.
c) Berechnen Sie das 95%-Konfidenzintervall für den durchschnittlichen Unterschied von APGAR-1-Werten und APGAR-5-Werten. Wie entwickeln sich Neugeborene in den ersten 5 Minuten nach der Geburt?
d) Wenn Neugeborene mit einem APGAR-1-Wert ≤ 3 in einem kritischen Zustand sind, berechnen Sie das 90%-Konfidenzintervall für den Unterschied der Anteile von Neugeborenen in kritischem Zustand zwischen Müttern, die *während* der Schwangerschaft geraucht haben und den Nichtraucherinnen.
e) Hat das Alter der Mutter einen signifikanten Einfluss auf den Anteil an Neugeborenen in kritischem Zustand?

Für die Lösung siehe Abschnitt 2.62

1.10 Signifikanztests

1.10.1 Aufgabe 1.10.1 Wirkstoffkonzentration

Die Wirkstoffkonzentration einer Zufallsstichprobe von 10 Arzneimittelbehältern aus einer Charge beträgt (in mg/mm^3)

17.6 19.2 21.3 15.1 17.6 18.9 16.2 18.3 19.0 16.4

a) Übertragen Sie die Daten in ein Datenframe mit der Variable `Konzentration`.
b) Testen Sie die zweiseitige Hypothese $H_0 : \mu = 18$ versus $H_1 : \mu \neq 18$ mit einem Signifikanzniveau von $\alpha = 0,05$.
c) Testen Sie die zweiseitige Hypothese $H_0 : \mu = 19,5$ versus $H_1 : \mu \neq 19,5$ mit den Signifikanzniveaus von $\alpha = 0,05$ und $0,01$. Wie beeinflusst das Signifikanzniveau das Testergebnis?
d) Testen Sie die zweiseitige Hypothese $H_0 : \mu = 17$ versus $H_1 : \mu \neq 17$ mit einem Signifikanzniveau von $\alpha = 0,05$. Testen Sie ebenfalls die Hypothesen $H_0 : \mu = 17$ versus $H_1 : \mu > 17$ mit $\alpha = 0,05$. Was ist der Unterschied zwischen den p-Werten des zweiseitigen und des einseitigen Tests?
e) Wenn der Hersteller angibt, die Konzentration des Wirkstoffs erhöht zu haben (im Vergleich zu früheren Chargen, bei denen der Mittelwert der Konzentration 17 mg/mm^3 war), können wir ihm glauben?
f) Welche Fallzahl würde benötigt, um einen Konzentrationsanstieg von 0,5 mg/mm^3 zu erkennen (mit $\alpha = 0,05$ und einer Power von $1 - \beta = 0,8$)?

[11] https://github.com/rkward-community/rk.Teaching, auch verfügbar unter https://www.produnis.de/R/data/neonates.RData

Für die Lösung siehe Abschnitt 2.63

1.10.2 Aufgabe 1.10.2 Bibliotheksnutzung

In einer von einer Universität durchgeführten Umfrage über die Nutzung der Bibliothek wurde eine Stichprobe von 34 Studierenden gefragt, ob sie mindestens einmal pro Woche in die Bibliothek gehen.

> nein ja nein nein nein ja nein ja ja ja ja nein ja nein ja nein nein nein ja ja ja nein
> nein ja nein nein ja ja nein nein ja nein ja nein

a) Übertragen Sie die Daten in ein Datenframe mit der Variable `bib`.
b) Testen Sie die Hypothese, dass der Anteil an Studierenden, die wöchentlich die Bibliothek nutzen, größer als 40% ist.

Für die Lösung siehe Abschnitt 2.64

1.10.3 Aufgabe 1.10.3 Laufen lernen

Eine Studie möchte untersuchen, ob Babies aus den unterschiedlichen Populationen A und B zu unterschiedlichen Zeiten anfangen zu laufen. In folgender Tabelle ist das Alter der Babies in Monaten aufgeführt, zu welchem sie mit dem Laufen anfingen.

A	9.5	10.5	9.0	9.8	10.0	13.0	10.0	13.5	10.0	9.8		
B	12.5	9.5	13.5	13.8	12.0	13.8	12.5	9.5	12.0	13.5	12.0	12.0

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `Alter` und `Population`.
b) Testen Sie die Hypothese, dass das durchschnittliche Alter in den Populationen unterschiedlich ist, mit $\alpha = 0,05$.

Für die Lösung siehe Abschnitt 2.65

1.10.4 Aufgabe 1.10.4 Bronchialretention

Forschende haben bei Rauchern einen größeren Atemwegswiderstand festgestellt als bei Nichtrauchern. Zur Überprüfung wurde bei 12 Probanden der Prozentsatz der tracheobronchialen Retention gemessen als sie Raucher waren und ein Jahr nach dem Rauchstopp.

Rauchen	Nichrauchen
60.6	47.5
12.0	13.3
56.0	33.0
75.2	55.2
12.5	21.9
29.7	27.9
57.2	54.3
62.7	13.9
28.7	8.90
66.0	46.1
25.2	29.8
40.1	36.2

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `vorher` und `nachher`.
b) Testen Sie, ob sich die Bronchialretention nach dem Rauchstopp verringert.

Für die Lösung siehe Abschnitt 2.66

1.10.5 Aufgabe 1.10.5 Prüfungen vormittags und nachmittags

In einem Kurs gibt es zwei Gruppen von Studierenden, eine am Vormittag und die andere am Nachmittag. Unter der Vormittagsgruppe haben 55 von 80 Studierenden bestanden, während in der Nachmittagsgruppe 32 von 90 Studierenden bestanden haben.

a) Gibt es signifikante Unterschiede zwischen den Prozentsätzen der Studiereden, die am Vormittag und am Nachmittag bestanden haben? Kann man daraus schließen, dass der Stundenplan die Ursache für diese Unterschiede ist?

Für die Lösung siehe Abschnitt 2.67

1.10.6 Aufgabe 1.10.6 Pulsmessung

Der Datensatz `pulse` von rk.Teaching[12] enthält Informationen über den Puls einer Stichprobe von Personen nach verschiedenen Übungen:

- Ruhepuls in Schlägen pro Minute (`pulse1`),
- Puls nach Bewegung in Schlägen pro Minute (`pulse2`),
- Art der Bewegung (`type`),
- Geschlecht (`sex`) und Gewicht (`weight`)

a) Testen Sie, ob der Ruhepuls weniger als 75 Schläge pro Minute beträgt.
b) Welcher Stichprobenumfang ist erforderlich, um einen Anstieg des Ruhepulses um 2 Schläge pro Minute mit einem Signifikanzniveau von 0,05 und einer Power von 0,9 festzustellen?
c) Testen Sie, ob der Puls nach dem Laufen größer als 85 Schläge pro Minute ist.
d) Eine Person hat eine leichte Tachykardie, wenn der Ruhepuls größer als 90 Schläge pro Minute ist. Prüfen Sie, ob der Prozentsatz der Personen mit leichter Tachykardie größer als 5% ist.
e) Kann man mit 95%iger Sicherheit schließen, dass Bewegung den Puls erhöht? Und bei einem Signifikanzniveau von $\alpha = 0,01$?
f) Gibt es einen Unterschied zwischen den durchschnittlichen Pulsschlägen nach dem Gehen und dem Laufen?
g) Gibt es einen Unterschied zwischen den Mittelwerten des Ruhepulses von Männern und Frauen? Und nach dem Laufen?

Für die Lösung siehe Abschnitt 2.68

[12] https://github.com/rkward-community/rk.Teaching, auch verfügbar unter https://www.produnis.de/R/data/pulse.RData

1.11 Varianzanalysen (ANOVA)

1.11.1 Aufgabe 1.11.1 Aknetherapie

In einer Studie wird versucht, die Wirksamkeit von drei Therapieprogrammen A, B und C zur Behandlung von von Akne zu bestimmen. Die Teilnehmer der Studie wurden nach dem Zufallsprinzip in drei Gruppen eingeteilt, und in jeder Gruppe wurde eine der Behandlungen durchgeführt. Nach 16 Wochen Behandlung wurde der prozentuale Rückgang der Akneläsionen gemessen.

Reduction in percentage of lesions

Treatment A		Treatment B		Treatment C	
48.6	50.8	68.0	71.9	67.5	61.4
49.4	47.1	67.0	71.5	62.5	67.4
50.1	52.5	70.1	69.9	64.2	65.4
49.8	49.0	64.5	68.9	62.5	63.2
50.6	46.7	68.0	67.8	63.9	61.2
		68.3	68.9	64.8	60.5
				62.3	

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `Therapie` und `Aknereduktion`.
b) Plotten Sie die Aknereduktion für jede Therapie. Sind Unterschiede erkennbar?
c) Führen Sie eine ANOVA durch. Gibt es signifikante Unterschiede zwischen den Therapien?
d) Berechnen Sie die Konfidenzintervalle für die paarweisen Unterschiede zwischen den drei Behandlungen. Bei welchen Behandlungen gibt es signifikante Unterschiede?
e) Plotten Sie diese Konfidenzintervalle.

Für die Lösung siehe Abschnitt 2.69

1.11.2 Aufgabe 1.11.2 Schulranking

Um zu prüfen, ob es zwischen den Schulen einer Stadt Unterschiede in den sportlichen Leistungen gibt, wurde eine Zufallsstichprobe von 8 Schülern jeder Schule gezogen. Die erreichten Punkte bei einem Sportwettkampf (von 1 bis 10) der jeweiligen Schüler sind in der folgenden Tabelle dargestellt.

A	B	C	D	E
5.5	6.1	4.9	3.2	6.7
5.2	7.2	5.5	3.3	5.8
5.9	5.5	6.1	5.5	5.4
7.1	6.7	6.1	5.7	5.5
6.2	7.6	6.2	6.0	4.9

5.9	5.9	6.4	6.1	6.2
5.3	8.1	6.9	4.7	6.1
6.2	8.3	4.5	5.1	7.0

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `Schule` und `Punkte`.
b) Plotten Sie die durchschnittlich erreichten Punkte pro Schule. Sind Unterschiede erkennbar?
c) Führen Sie eine ANOVA durch. Gibt es signifikante Unterschiede zwischen den Schulen?
d) In welcher Schule sind die sportlichen Leistungen am besten?

Für die Lösung siehe Abschnitt 2.70

1.11.3 Aufgabe 1.11.3 Puls und Herzkrankheit

Die nachstehende Tabelle zeigt den Puls (in Schlägen pro Minute) von vier Patientengruppen: Kontrollen (`A`), Patienten mit Angina pectoris (`B`), Patienten mit Herzrhythmusstörungen (`C`) und Patienten, die sich von einem Herzinfarkt erholt haben (`D`).

A	*B*	*C*	*D*
83	81	75	61
61	65	68	75
80	77	80	78
63	87	80	80
67	95	74	68
89	89	78	65
71	103	69	68
73	89	72	69
70	78	76	70
66	83	75	79
57	91	69	61

a) Gibt es laut den Daten signifikante Unterschiede zwischen den vier Gruppen?

Für die Lösung siehe Abschnitt 2.71

1.11.4 Aufgabe 1.11.4 Kohlenmonoxid

Die folgende Tabelle zeigt die Atemfrequenz (Atemzüge pro Minute) bei einer Stichprobe von Laborratten, die drei Konzentrationen von Kohlenmonoxid ausgesetzt waren.

Low	Medium	High
36	43	45
33	38	39
35	41	33
39	34	39
41	28	33
41	44	26
44	30	39
45	31	29

a) Gibt es laut den Daten signifikante Unterschiede zwischen den drei Gruppen?

Für die Lösung siehe Abschnitt 2.72

1.12 Chiquadratests für Anteilswerte

1.12.1 Aufgabe 1.12.1 Magengeschwür

Die folgende Tabelle enthält die Blutgruppe einer Stichprobe von 1655 Patienten mit Magengeschwüren und 10000 Patienten ohne Magengeschwüre Patienten.

	0	A	B	AB
Geschwür	911	579	124	41
kein Geschwür	4578	4219	890	313

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `Geschwuer` und `Blutgruppe`.
b) Führen Sie einen Chiquadrattest auf die Hypothese durch, dass die Geschwüre von der Blutgruppe abhängig sind.
c) Gibt es in Anbetracht der Ergebnisse des Vergleichs einen Zusammenhang zwischen dem Magengeschwür und der Blutgruppe? Können wir behaupten, dass der Anteil der Ulkuspatienten je nach Blutgruppe unterschiedlich ist?

Für die Lösung siehe Abschnitt 2.73

1.12.2 Aufgabe 1.12.2 Blutgruppen

Mitchell et al. (1976) untersuchten die Verteilung der Blutgruppen in einer Stichprobe von 478 Personen aus verschiedenen Regionen im Südwesten Schottlands. Sie erhielten die folgenden Ergebnisse:

	Eskdale	Annandale	Nithsdale	
A	33	54	98	185
B	6	14	35	55
O	56	52	115	223
AB	5	5	5	15
	100	125	253	478

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `Region` und `Blutgruppe`.
b) Führen Sie einen Chiquadrattest auf die Hypothese durch, dass die Blutgruppe von der Region abhängig sind.
c) Gibt es in Anbetracht der Ergebnisse einen Zusammenhang zwischen der Blutgruppe und der Region? Können wir behaupten, dass die Region keinen Einfluss auf die Blutgruppe hat?

Für die Lösung siehe Abschnitt 2.74

1.12.3 Aufgabe 1.12.3 Rauchen und Geschlecht

In einer Studie wurd versucht festzustellen, ob das Rauchen mit dem Geschlecht zusammenhängt. Es wurden 9 Männer und 17 Frauen befragt. Unter den männlichen Probanden gab es 2 Raucher, während in der weiblichen Stichprobe 6 Raucherinnen waren.

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `Rauchen` und `Geschlecht`.
b) Führen Sie einen Chi-Quadrat-Test durch, um festzustellen, ob das Rauchen mit dem Geschlecht zusammenhängt.
c) Ist die Verteilung der Raucher bei beiden Geschlechtern gleich?

Für die Lösung siehe Abschnitt 2.75

1.12.4 Aufgabe 1.12.4 Migräne

Um die Wirksamkeit von zwei Medikamenten gegen Migräne zu vergleichen, wurden 20 Personen, die häufig unter Migräne litten, ausgewählt und die beiden Medikamente zu verschiedenen Zeitpunkten ausprobiert. Die folgende Tabelle zeigt die Anzahl der Personen, die eine gewisse Linderung erfuhren.

Drug 1	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes
Drug 2	No	No	Yes	No	Yes	Yes	No	No	No	No

Drug 1	Yes	No	Yes	No	Yes	Yes	Yes	No	Yes	Yes
Drug 2	Yes	No	Yes	No	No	Yes	No	Yes	No	No

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `drug1` und `drug2`.
b) Führen Sie einen McNemar-Test durch, um festzustellen, ob die Linderung mit dem Medikament zusammenhängt.
c) Können wir nach dem Ergebnis des Tests behaupten, dass die Linderung der Migräne vom Medikament abhängt? Wenn ja, welches Medikament bewirkt eine signifikant höhere Linderung?

Für die Lösung siehe Abschnitt 2.76

1.12.5 Aufgabe 1.12.5 Komatös

Eine Studie versucht zu bestimmen, ob Patienten, die bei der Ankunft im Krankenhaus komatös sind, eine schlechtere Prognose (Überleben oder Sterben) haben.

	nicht komatös	komatös	
überleben	484	37	521
verstorben	118	89	207
	602	126	728

a) Ist ein komatöser Zustand bei der Ankunft im Krankenhaus ein Risikofaktor zu versterben?

Für die Lösung siehe Abschnitt 2.77

1.12.6 Aufgabe 1.12.6 Heilung

Die Heilung einer Krankheit, die durch zwei Behandlungen A und B hervorgerufen wird, wird in drei Kategorien eingeteilt: sehr gut, gut und schlecht.
Die Behandlung A wird bei 32 Patienten angewandt und B bei 28. Bei Medikament A konnten 10 von insgesamt 22 `sehr guten` Heilungen, 14 von insgesamt 24 `guten` Heilungen und 8 von insgesamt 14 `schlechten` Heilungen beobachtet werden. Ist die Wirksamkeit der beiden Behandlungen die gleiche?

Für die Lösung siehe Abschnitt 2.78

1.12.7 Aufgabe 1.12.7 Facherfolg

Um festzustellen, ob Frauen in einem Fach erfolgreicher sind als Männer, wurde eine Stichprobe von 10 Frauen und 10 Männern gezogen. Beide Gruppen wurden von einem Lehrer geprüft, der immer 40% der Prüflinge durchfallen lässt. Wenn man weiß, dass nur 2 Männer bestanden haben, können wir dann behaupten, dass Frauen in diesem Fach erfolgreicher sind als Männer?

Für die Lösung siehe Abschnitt 2.79

1.12.8 Aufgabe 1.12.8 Statistikdozent

150 Studierende wurden befragt, ob ihnen die Lehrmethoden von zwei Biostatistik-Dozenten (Hans und Erna) gefallen. Die Ergebnisse sind in der nachstehenden Tabelle aufgeführt:

	like Hans	dislike Hans
like Erna	37	48
dislike Erna	44	21

Können wir bestätigen, dass es unterschiedliche Meinungen über Hans und Erna gibt?

Für die Lösung siehe Abschnitt 2.80

2 Lösungen

Die hier vorgestellten Lösungen stellen immer nur *eine mögliche* Vorgehensweisen dar und sind sicherlich nicht der Weisheit letzter Schluss. In R führen viele Wege nach Rom, und wenn Sie mit anderem Code zu den richtigen Ergebnissen kommen, dann ist das völlig in Ordnung.

2.1 Lösung zur Aufgabe 1.1.1

a) Erstellen Sie ein Datenframe mit der Variable `Kinder` und übertragen Sie die Daten.

```
# erzeuge Datenframe
df <- data.frame(Kinder = c(1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0,
                            2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2))
```

b) Erzeugen Sie eine einfache Häufigkeitstabelle

```
# erzeuge Datenframe
table(df$Kinder)
```

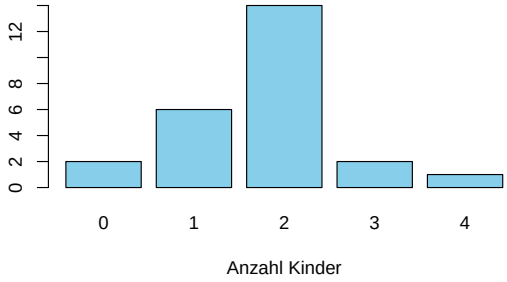
```
 0  1  2  3  4
 2  6 14  2  1
```

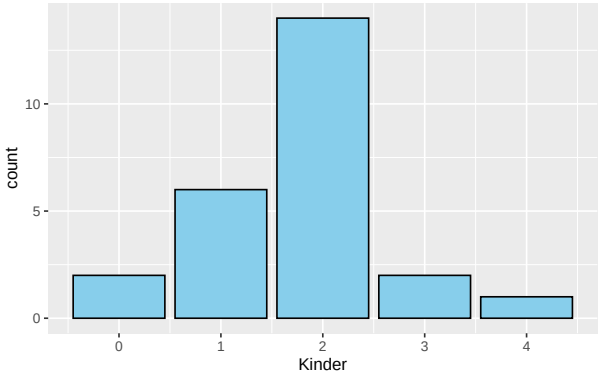
```
# oder
xtabs(~Kinder, data=df)
```

```
Kinder
 0  1  2  3  4
 2  6 14  2  1
```

c) Erzeugen Sie ein Balkendiagramm der Häufigkeiten

```
# Balkendiagramm mit R-Base
barplot(table(df$Kinder), col="skyblue", xlab="Anzahl Kinder")
# mit ggplot
ggplot(df, aes(x=Kinder)) +
          geom_bar(fill="skyblue", color="black")
```





d) Erzeugen Sie eine vollständige Häufigkeitstabelle, inklusive absoluter, relativer und jeweils kumulativer Häufigkeiten

```
## zu Fuß
# kumulierte absoulte Häufigkeiten
cumsum(table(df$Kinder))
```

```
 0  1  2  3  4
 2  8 22 24 25
```

```
# relative Häufigkeiten
(table(df$Kinder)/length(df$Kinder))*100
```

```
 0  1  2  3  4
 8 24 56  8  4
```

```
# kumulierterelative Häufigkeiten
cumsum(table(df$Kinder)/length(df$Kinder))*100
```

```
  0   1   2   3   4
  8  32  88  96 100
```

```
## einfacher
# erzeuge vollständige Häufigkeitstabelle
jgsbook::freqTable(df$Kinder)
```

Wert	Haeufig	Hkum	Relativ	Rkum
0	2	2	8	8
1	6	8	24	32
2	14	22	56	88
3	2	24	8	96
4	1	25	4	100

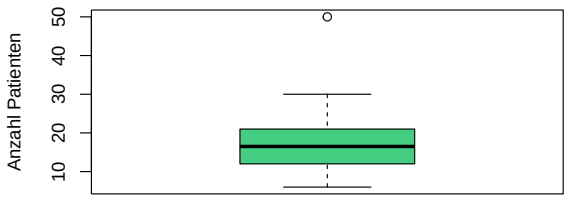
2.2 Lösung zur Aufgabe 1.1.2

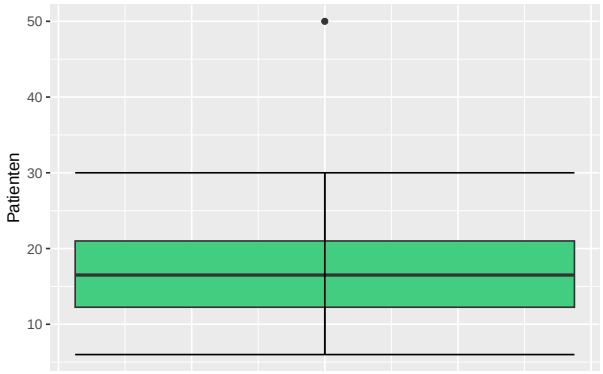
a) Erstellen Sie ein Datenframe mit der Variable `Patienten` und übertragen Sie die Daten.

```
# erzeuge Datenframe
df <- data.frame(Patienten = c(15, 23, 12, 10, 28, 50, 12, 17, 20,
                               21, 18, 13, 11, 12, 26, 30,  6, 16,
                               19, 22, 14, 17, 21, 28,  9, 16, 13,
                               11, 16, 20))
```

b) Erzeugen Sie ein Boxplot und entfernen Sie etwaige Ausreißer.

```
# Boxplot mit Rbase
boxplot(df$Patienten, col="seagreen3", ylab="Anzahl Patienten")
# Boxplot mit ggplot
ggplot(df, aes(y=Patienten)) +
    geom_boxplot(fill="seagreen3") +
    # whiskers
    stat_boxplot(geom="errorbar") +
    theme(axis.ticks.x=element_blank(),
          axis.text.x=element_blank())
```





Es ist ein Ausreißer enthalten.

```
# entferne Ausreißer für weiteres Vorgehen
df <- subset(df, Patienten < 50)
```

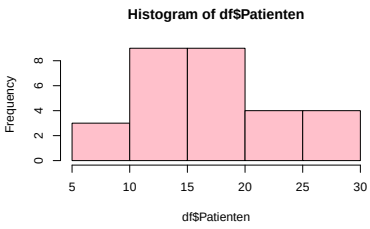
c) Erzeugen Sie eine Häufigkeitstabelle, welche die Daten in 5 Klassen gruppiert.

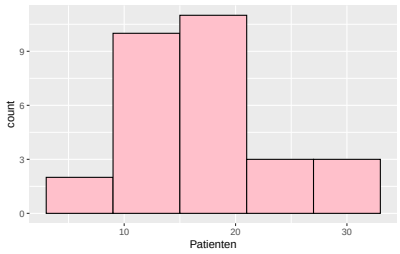
```
# klassiere in 5 Gruppen
gruppen <- cut(df$Patienten, breaks = 5, ordered_result = TRUE)
# Häufigkeitstabelle
table(gruppen)
```

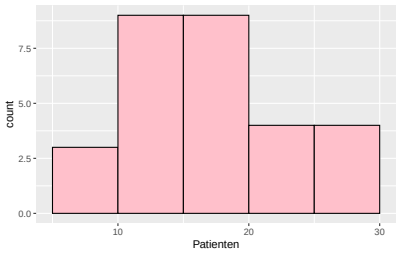
```
gruppen
(5.98,10.8] (10.8,15.6] (15.6,20.4] (20.4,25.2]   (25.2,30]
          3           9           9           4           4
```

d) Erzeugen Sie ein Histogram der klassierten absoluten Häufigkeiten.

```
# Histogram mit Rbase
hist(df$Patienten, col="pink")
# mit ggplot werden andere Breaks erzeugt
ggplot(df, aes(x=Patienten)) +
   geom_histogram(fill="pink", color="black",
                  bins=5)
# also die Klassengrenzen manuell festlegen
ggplot(df, aes(x=Patienten)) +
   geom_histogram(fill="pink", color="black",
                  breaks=c(5, 10, 15, 20, 25, 30))
```



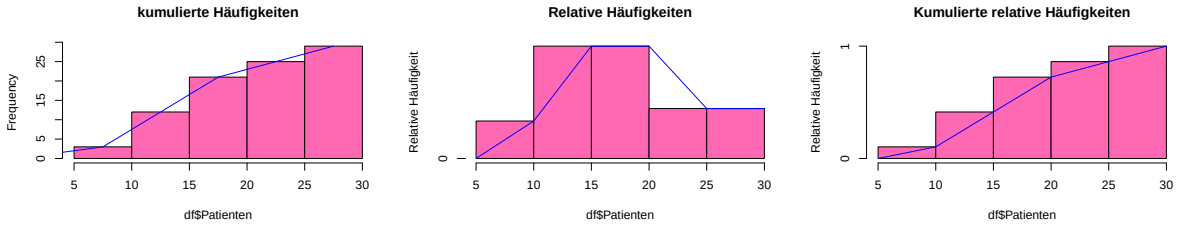




e) Erzeugen Sie ebenso Histogramme der relativen und jeweils kumulativen Häufigkeiten, inklusive Polygonzügen.

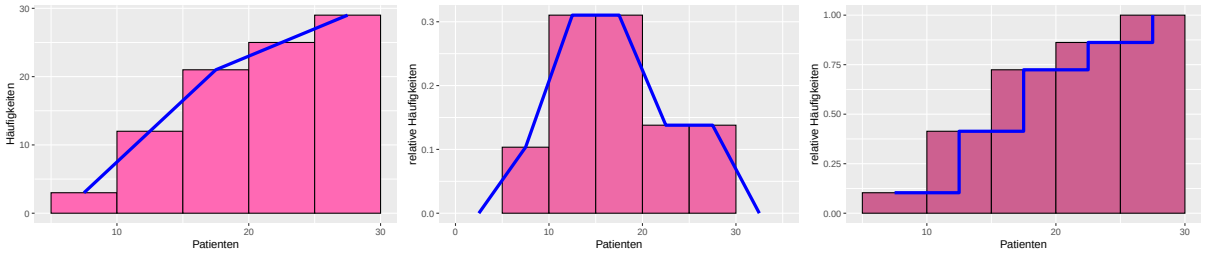
Mit R base können wir wie folgt vorgehen.

```r
# 1. kumulierte absolute Häufigkeiten
#-------------------------------------
# speichere Histogramm in Objekt h
h <- hist(df$Patienten, plot=FALSE)

# ersetze die Zellen durch kumulierte Häufigkeiten
h$counts <- cumsum(h$counts)

# plotte das kumulative Histogram
plot(h, col="hotpink", main = "kumulierte Häufigkeiten")
# füge Polygonzug hinzu
lines(c(0, h$mids),c(0,h$counts), col="blue") # type="s"
##--------------------------------------#


# 2. Histogram der relativen Häufigkeiten #
##--------------------------------------#
# speichere Histogramm in Objekt h
h <- hist(df$Patienten, plot=FALSE)

# relative Häufigkeiten
h$counts <- h$counts/sum(h$counts)
### plot
plot(h, col="hotpink", main = "Relative Häufigkeiten",
     ylab = "Relative Häufigkeit" )
# Polygon hinzufügen
lines(h$breaks, c(0, h$counts), col = "blue") # add type="s" if you like
##-----------------------------------------------------#


# 3. Histogram der kumulierten relativen Häufigkeiten #
##-----------------------------------------------------#
# speichere Histogramm in Objekt h
h <- hist(df$Patienten, plot=FALSE)

# kumulative relative Häufigkeiten
h$counts <- cumsum(h$counts)/sum(h$counts)
### plot
plot(h, col="hotpink", main = "Kumulierte relative Häufigkeiten",
     ylab = "Relative Häufigkeit" )
# Polygon hinzufügen
lines(h$breaks, c(0, h$counts), col = "blue") # add type="s" if you like
```



Im Tidyverse können wir so vorgehen.

```
### Mittels ggplot()
# Klassengrenzen festlegen
breaks = c(5, 10, 15, 20, 25, 30)

# kumulierte Häufigkeiten
ggplot(df, aes(x=Patienten)) +
  ylab("Häufigkeiten")+
  geom_histogram(aes(y=cumsum(after_stat(count))),
                 fill="hotpink", color="black",
                 breaks=breaks) +
  stat_bin(aes(y=cumsum(after_stat(count))),
           breaks=breaks,
           geom="line", color="blue", linewidth=1.5) # oder geom="step"
#-----------------

# relative Häufigkeiten
ggplot(df, aes(x=Patienten))+
  ylab("relative Häufigkeiten")+
  geom_histogram(aes(y=after_stat(count)/sum(after_stat(count))),
                 breaks=breaks, fill="hotpink2", color="black") +
  geom_freqpoly(aes(y=after_stat(count)/sum(after_stat(count))),
                 breaks=breaks, color="blue", linewidth=1.5)
#-----------------

# kumulierte relative Häufigkeiten
ggplot(df, aes(x=Patienten)) +
  ylab("relative Häufigkeiten")+
  geom_histogram(aes(y=cumsum(after_stat(count)/sum(after_stat(count)))),
                 breaks=breaks, fill="hotpink3", color="black") +
  stat_bin(aes(y=cumsum(after_stat(count)/sum(after_stat(count)))),
                 breaks=breaks,
                 geom="step", color="blue", linewidth=1.5) # oder geom="line"
```



2.3 Lösung zur Aufgabe 1.1.3

a) Erstellen Sie ein Datenframe mit der Variable `Blutgruppe` und übertragen Sie die Daten.

```
# Übertrage Daten
df <- data.frame(Blutgruppe = factor(c("A", "B", "B", "A", "AB", "0", "0", "A",
                                       "B", "B", "A", "A", "A", "A", "AB", "A",
                                       "A", "A", "B", "0", "B", "B", "B", "A",
                                       "A", "A", "0", "A", "AB", "0")))
```

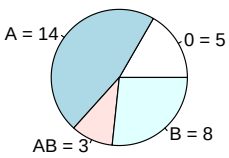
b) Erzeugen Sie eine Häufigkeitstabelle

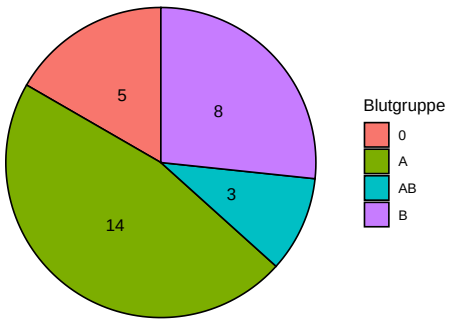
```
table(df$Blutgruppe)
```

```
 0  A AB  B
 5 14  3  8
```

c) Erzeugen Sie ein Kreisdiagramm

```
# mit R base
pie(table(df$Blutgruppe),
    labels = paste(levels(df$Blutgruppe),"=",
                   as.numeric(table(df$Blutgruppe))
                  )
    )
#------

# mit ggplot benötigen wir ein Hilfsdatenframe
df2 <- as.data.frame(table(df$Blutgruppe))
colnames(df2) <- c("Blutgruppe", "Wert")

ggplot(df2, aes(x="", y=Wert, fill=Blutgruppe)) +
  geom_col(color="black") +
  # Werte schreiben
  geom_text(aes(label = Wert),
            position = position_stack(vjust = 0.5)) +
  # verbiege zu Kreisdiagramm
  coord_polar(theta="y") +
  # entferne Achsen und Ticks
  theme_void()
```





2.4 Lösung zur Aufgabe 1.1.4

a) Erstellen Sie ein Datenframe mit den Variablen `Alter` und `Familienstand` und übertragen Sie die Daten.

```
df <- data.frame(Alter = c(31, 45, 35, 65, 21, 38, 62, 22, 31,
                           72, 39, 62, 59, 25, 44, 54,
                           80, 68, 65, 40, 78, 69, 75,
                           31, 65, 59, 58, 50),
                 Familienstand = c( rep("Single", 9),
                                    rep("Verheiratet", 7),
                                    rep("Verwitwet", 7),
                                    rep("Geschieden", 5)
                                  )
                 )
```

b) Erzeugen Sie für jeden `Familienstand` eine Häufigkeitstabelle des `Alters`.

```
# Singles
df2 <- subset(df, Familienstand=="Single")
table(df2$Alter)
```

```
21 22 31 35 38 45 62 65
 1  1  2  1  1  1  1  1
```

```
# Verheiratet
df2 <- subset(df, Familienstand=="Verheiratet")
table(df2$Alter)
```

```
25 39 44 54 59 62 72
 1  1  1  1  1  1  1
```

```
# Verwitwet
df2 <- subset(df, Familienstand=="Verwitwet")
table(df2$Alter)
```

```
40 65 68 69 75 78 80
 1  1  1  1  1  1  1
```

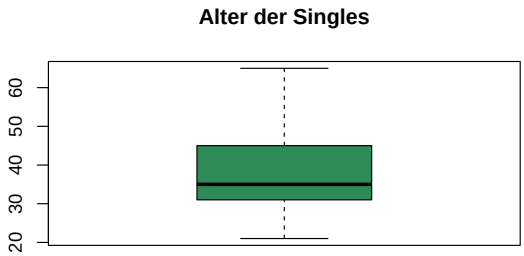
```
# Geschieden
df2 <- subset(df, Familienstand=="Geschieden")
table(df2$Alter)
```

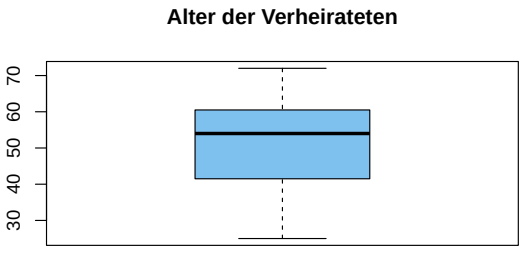
```
31 50 58 59 65
 1  1  1  1  1
```

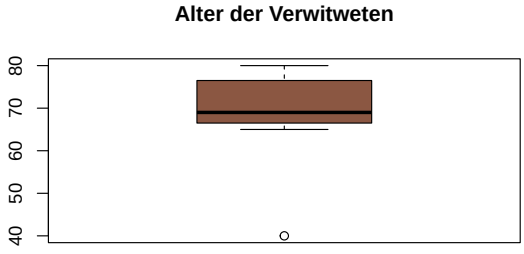
c) Erzeugen Sie für jeden `Familienstand` eine Boxplot des `Alters`. Gibt es Ausreißer? In welcher Gruppe streut das Alter am meisten?

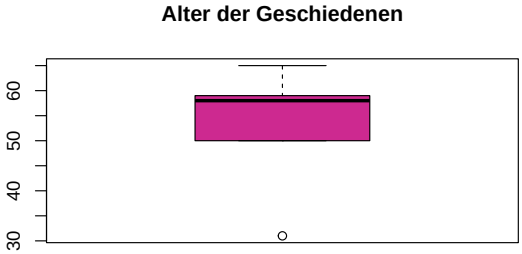
Mit R base können wir so vorgehen.

```
# Singles
df2 <- subset(df, Familienstand=="Single")
boxplot(df2$Alter, main="Alter der Singles", col="seagreen")
#---------------------------------------

# Verheiratet
df2 <- subset(df, Familienstand=="Verheiratet")
boxplot(df2$Alter, main="Alter der Verheirateten", col="skyblue2")
#---------------------------------------

# Verwitwet
df2 <- subset(df, Familienstand=="Verwitwet")
boxplot(df2$Alter, main="Alter der Verwitweten", col="lightsalmon4")
#---------------------------------------

# Geschieden
df2 <- subset(df, Familienstand=="Geschieden")
boxplot(df2$Alter, main="Alter der Geschiedenen", col="maroon3")
#---------------------------------------
```



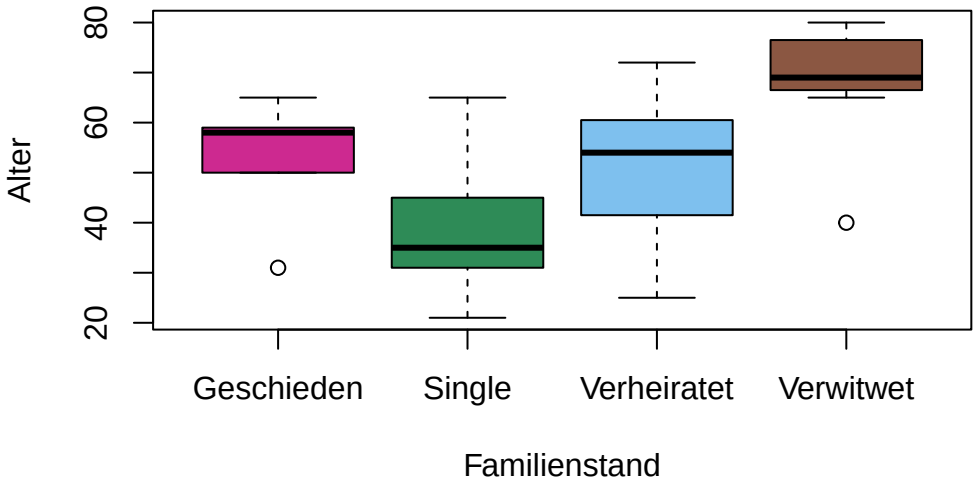






Oder alle auf einmal:

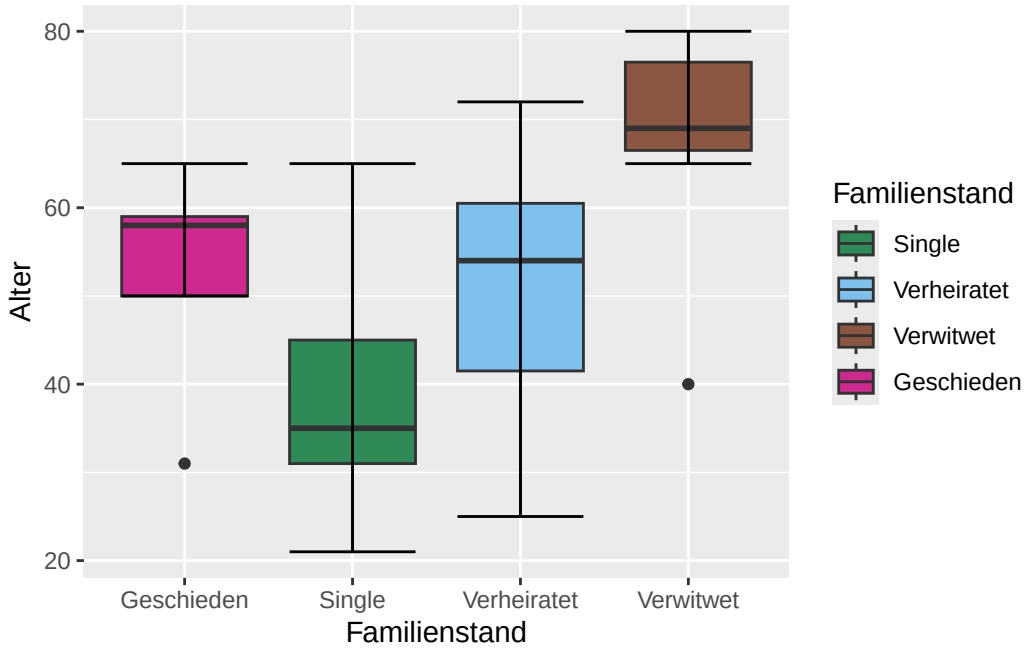
```
boxplot(Alter ~ Familienstand , data=df,
        col=c("maroon3", "seagreen", "skyblue2", "lightsalmon4"))
```



Es sind Ausreißer erkennbar in der Gruppe der Verwitweten und der Geschiedenen.

Im Tidyverse können wir so vorgehen:

```
ggplot(df, aes(y=Alter, x=Familienstand)) +
  geom_boxplot(aes(fill=Familienstand)) +
  stat_boxplot(geom="errorbar")+
  scale_fill_manual(values=c("seagreen", "skyblue2",
                             "lightsalmon4", "maroon3"),
                    breaks=c("Single", "Verheiratet",
                             "Verwitwet", "Geschieden"))
```

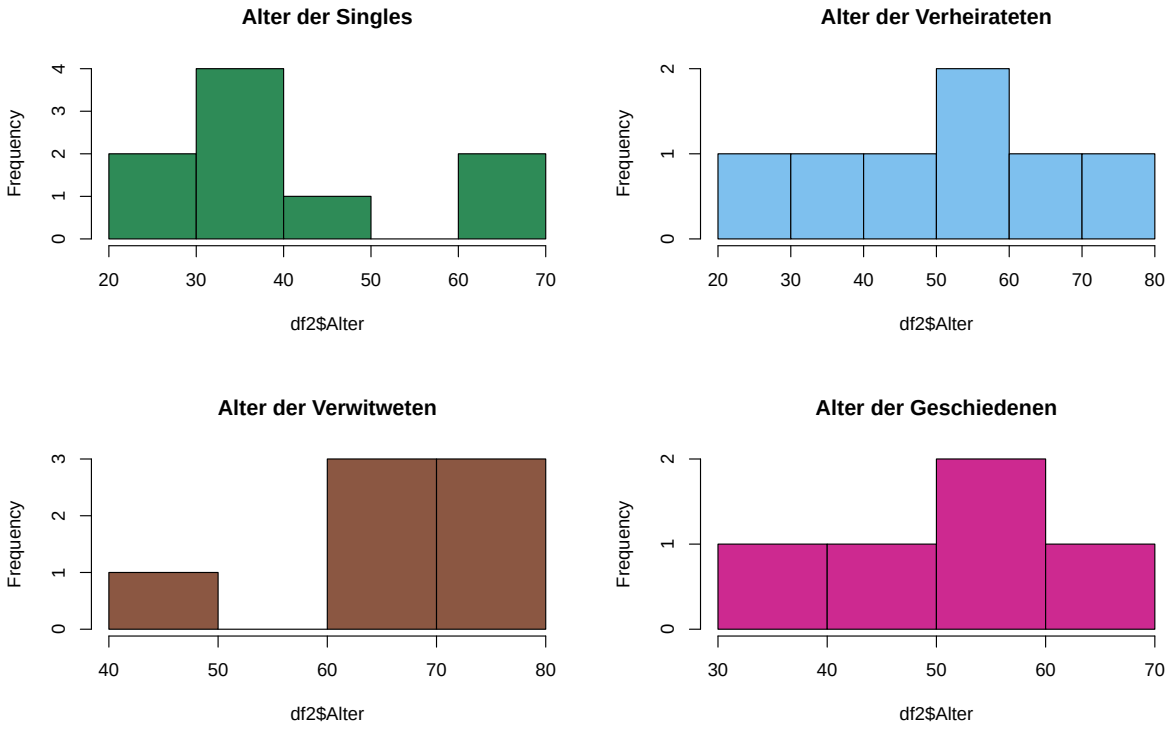


Auch hier sind die Ausreißer in den Gruppen der Verwitweten und der Geschiedenen erkennbar.

d) Erzeugen Sie für jeden `Familienstand` eine Histogram des `Alters`. Wie unterscheiden sich die Histogramme?

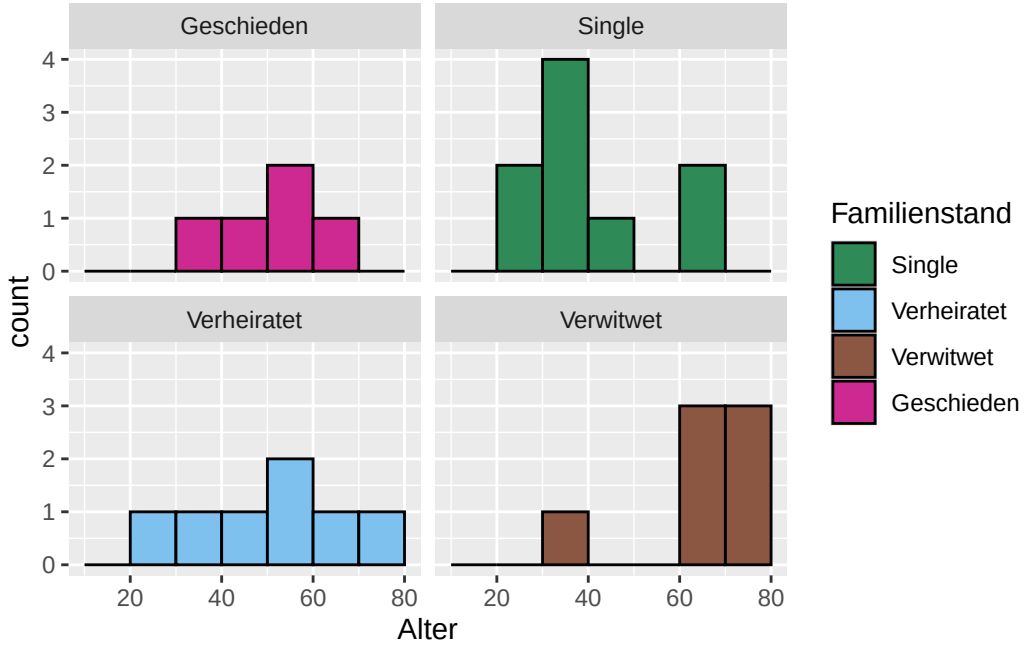
Mit R base können wir so vorgehen

```
# Singles
df2 <- subset(df, Familienstand=="Single")
hist(df2$Alter, main="Alter der Singles", col="seagreen")
#---------------------------------------

# Verheiratet
df2 <- subset(df, Familienstand=="Verheiratet")
hist(df2$Alter, main="Alter der Verheirateten", col="skyblue2")
#---------------------------------------

# Verwitwet
df2 <- subset(df, Familienstand=="Verwitwet")
hist(df2$Alter, main="Alter der Verwitweten", col="lightsalmon4")
#---------------------------------------

# Geschieden
df2 <- subset(df, Familienstand=="Geschieden")
hist(df2$Alter, main="Alter der Geschiedenen", col="maroon3")
#---------------------------------------
```



Im Tidyverse können wir so vorgehen:

```
breaks = c(seq(10,80,10))
ggplot(df, aes(x=Alter, fill=Familienstand)) +
    geom_histogram(breaks=breaks, color="black")+
    scale_fill_manual(values=c("seagreen", "skyblue2",
                               "lightsalmon4", "maroon3"),
                      breaks=c("Single", "Verheiratet",
                               "Verwitwet", "Geschieden"))+
    facet_wrap(~Familienstand)
```



2.5 Lösung zur Aufgabe 1.1.5

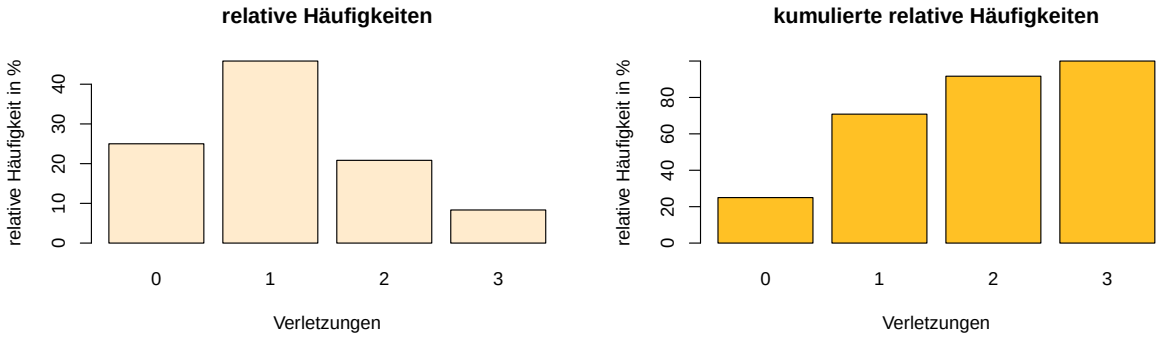
a) Erstellen Sie eine Häufigkeitstabelle

```
# erzeuge Daten
Verletzung <- c(0, 1, 2, 1, 3, 0, 1, 0, 1, 2, 0, 1, 1, 1, 2, 0,
                1, 3, 2, 1, 2, 1, 0, 1)

# Häufigkeitstabelle
xtabs(~Verletzung)
```

```
Verletzung
 0  1  2  3
 6 11  5  2
```

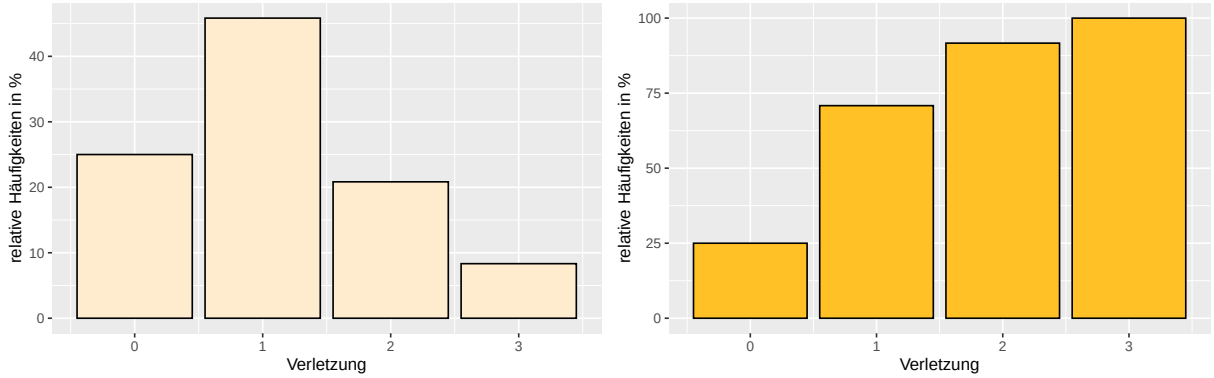
b) Erzeugen Sie ein Säulendiagramm der relativen und kumulativen relativen Häufigkeiten.

```
# relative Häufigkeiten als Barplot
barplot( table(Verletzung) / length(Verletzung)*100,
        col="blanchedalmond", main="relative Häufigkeiten",
        ylab="relative Häufigkeit in %", xlab="Verletzungen")
#-----------------------------------------------------

# relative kumulierte Häufigkeiten als Barplot
barplot( cumsum(table(Verletzung)) / sum(table(Verletzung))*100,
        col="goldenrod1", main="kumulierte relative Häufigkeiten",
        ylab="relative Häufigkeit in %", xlab="Verletzungen")
```



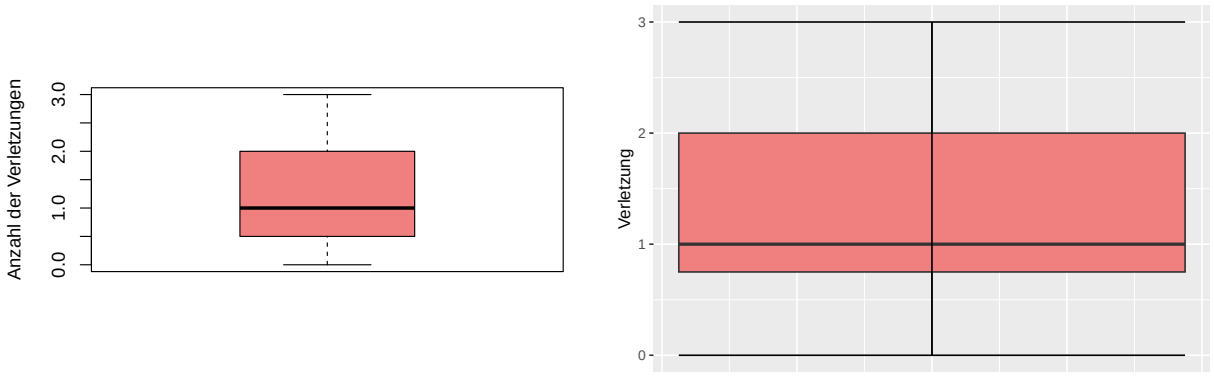
Mit ggplot

```
df <- data.frame(Verletzung)
ggplot(df, aes(x=Verletzung))+
    geom_bar(aes(y=after_stat(count)/sum(after_stat(count))*100),
            fill="blanchedalmond", color="black") +
    ylab("relative Häufigkeiten in %")
#-----------------------------

# kumulierte relative Häufigkeiten
ggplot(df, aes(x=Verletzung))+
    ylab("relative Häufigkeiten in %")+
    geom_bar(aes(y=cumsum(after_stat(count)/sum(after_stat(count)))*100),
            fill="goldenrod1", color="black")
```



c) Erzeugen Sie ein Boxplot

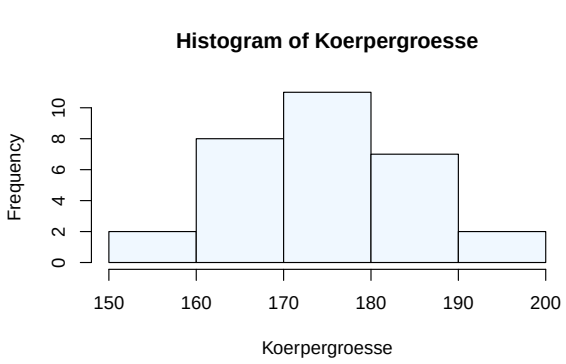
```
boxplot(Verletzung, col="lightcoral", ylab="Anzahl der Verletzungen")
#--------------------

# mit ggplot
ggplot(df, aes(y=Verletzung)) +
    geom_boxplot(fill="lightcoral") +
    stat_boxplot(geom="errorbar") +
    theme(axis.ticks.x=element_blank(),
          axis.text.x=element_blank())
```

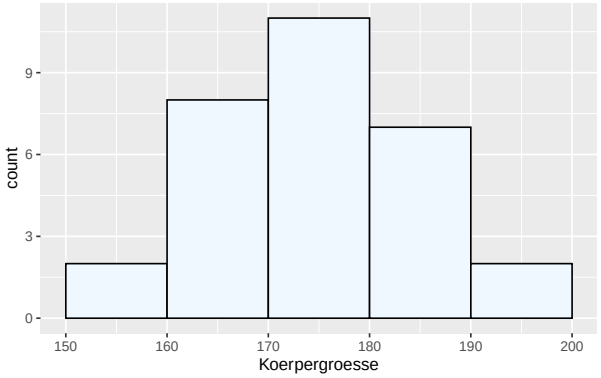


2.6 Lösung zur Aufgabe 1.1.6

a) Erstellen Sie ein Histogram der Körpergröße mit Klassen von 150cm bis 200cm, die jeweils 10cm breit sind.

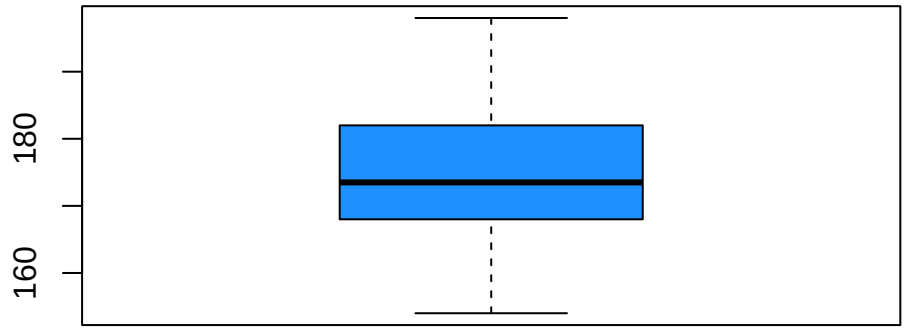
```
Koerpergroesse <- c(179, 173, 181, 170, 158, 174, 172, 166, 194, 185,
                    162, 187, 198, 177, 178, 165, 154, 188, 166, 171,
                    175, 182, 167, 169, 172, 186, 172, 176, 168, 187)

hist(Koerpergroesse, breaks=seq(150, 200, 10),
                     col="aliceblue")
#-----------------------

# mit ggplot
as.data.frame(Koerpergroesse) %>%
    ggplot(aes(x=Koerpergroesse)) +
        geom_histogram(breaks=seq(150, 200, 10),
                       fill="aliceblue", color="black")
```





b) Gibt es Ausreißer?

```
boxplot(Koerpergroesse, col="dodgerblue1")
```



Es sind keine Ausreißer erkennbar.

2.7 Lösung zur Aufgabe 1.1.7

```
# lade Daten
load("https://www.produnis.de/R/data/neonates.RData")
```

```
# Datensatz anschauen
summary(neonates)
```

```
     weight          gender                   age       smoke       cigarettes
 Min.   :2.021   male  :157   greater than 20:218   No :220   Min.   : 0.000
 1st Qu.:2.794   female:163   less than 20   :102   Yes:100   1st Qu.: 0.000
 Median :3.030                                               Median : 0.000
 Mean   :3.026                                               Mean   : 3.891
 3rd Qu.:3.267                                               3rd Qu.: 8.250
 Max.   :4.182                                               Max.   :22.000
 smoke.before     apgar1          apgar5
 No :185      Min.   :2.000   Min.   : 2.000
 Yes:135      1st Qu.:5.000   1st Qu.: 5.000
              Median :6.000   Median : 6.000
              Mean   :5.628   Mean   : 6.213
              3rd Qu.:6.000   3rd Qu.: 7.000
              Max.   :9.000   Max.   :10.000
```

a) Erstellen Sie die Häufigkeitstabelle des APGAR-Scores nach 1 Minute. Wenn ein Score von 3 oder weniger anzeigt, dass das Neugeborene in einem kritischen Zusatand ist, wie viel Prozent der Neugeborenen in der Stichprobe sind dann in einem kritischen Zustand?

```
# neue Variable "kritisch"
neonates$kritisch <- FALSE
# nur solche mit APGAR<4 sind kritisch
neonates$kritisch[neonates$apgar1<4] <- TRUE
# relative Häufigkeiten
table(neonates$kritisch) / length(neonates$kritisch) * 100
```

```
  FALSE    TRUE
92.1875  7.8125
```

7,81% der Neugeborenen sind in einem kritischen Zustand.

b) Erstellen Sie die Häufigkeitstabelle des Geburtsgewichts der Neugeborenen, indem Sie die Daten in Klassen mit einer Breite von 0,5 kg von 2 bis 4,5 kg einteilen. Welches Intervall enthält die meisten Neugeborenen?

```
# neue Variable für die Gewichtsklassifikation
neonates$gewiKat <- cut(neonates$weight, breaks=seq(2, 4.5, 0.5),
                        ordered_results=TRUE)
# einfache Häufigkeitstabelle
table(neonates$gewiKat)
```

```
(2,2.5] (2.5,3] (3,3.5] (3.5,4] (4,4.5]
     22     127     146      24       1
```

```
# oder vollständige
jgsbook::freqTable(neonates$gewiKat)
```

Wert	Haeufig	Hkum	Relativ	Rkum
(2,2.5]	22	22	6.88	6.88
(2.5,3]	127	149	39.69	46.57
(3,3.5]	146	295	45.62	92.19
(3.5,4]	24	319	7.50	99.69
(4,4.5]	1	320	0.31	100.00

Das Intervall von 3-3,5kg enthält die meisten Neugeborenen.

c) Vergleichen Sie die Häufigkeitsverteilung des APGAR-Scores nach 1 Minute für Mütter unter 20 Jahren und für Mütter über 20 Jahren. Welche Gruppe hat mehr Neugeborene in kritischem Zustand?

```
# gruppieren
gruppe1 <- neonates$apgar1[neonates$age=="less than 20"]
gruppe2 <- neonates$apgar1[neonates$age=="greater than 20"]

# Jünger als 20
jgsbook::freqTable(gruppe1)
```

Wert	Haeufig	Hkum	Relativ	Rkum
2	2	2	1.96	1.96
3	11	13	10.78	12.74
4	16	29	15.69	28.43
5	28	57	27.45	55.88
6	28	85	27.45	83.33
7	12	97	11.76	95.09
8	4	101	3.92	99.01
9	1	102	0.98	99.99

```
# Älter als 20
jgsbook::freqTable(gruppe2)
```

Wert	Haeufig	Hkum	Relativ	Rkum
2	2	2	0.92	0.92
3	10	12	4.59	5.51
4	22	34	10.09	15.60
5	53	87	24.31	39.91
6	69	156	31.65	71.56
7	34	190	15.60	87.16
8	24	214	11.01	98.17
9	4	218	1.83	100.00

In der Gruppe der unter-20-jährigen liegt der Prozentsatz an Neugeborenen mit APGAR-Werten kleinergleich 3 bei 12,74%. In der Gruppe der über-20-jährigen liegt der Prozentwert bei 5,51%. Es tritt also in der Gruppe der jüngeren Mütter häufiger auf.

d) Vergleichen Sie die relative Häufigkeitsverteilung des Geburtsgewichts der Neugeborenen, je nachdem, ob die Mutter während der Schwangerschaft geraucht hat oder nicht. Wenn ein Gewicht unter 2,5 kg als niedriges Gewicht gilt, welche Gruppe hat einen höheren Prozentsatz an Neugeborenen mit niedrigem Gewicht?

```
# (hatten wir oben schon)
# Gewichtsklassifikation
neonates$gewiKat <- cut(neonates$weight, breaks=seq(2, 4.5, 0.5),
                        right=FALSE, ordered_results=TRUE)
gruppe1 <- neonates$gewiKat[neonates$smoke=="No"]
gruppe2 <- neonates$gewiKat[neonates$smoke=="Yes"]

# Nichtraucherinnen
jgsbook::freqTable(gruppe1)
```

Wert	Haeufig	Hkum	Relativ	Rkum
[2,2.5)	5	5	2.27	2.27
[2.5,3)	75	80	34.09	36.36
[3,3.5)	119	199	54.09	90.45
[3.5,4)	20	219	9.09	99.54
[4,4.5)	1	220	0.45	99.99

```
# Raucherinnen
jgsbook::freqTable(gruppe2)
```

Wert	Haeufig	Hkum	Relativ	Rkum
[2,2.5)	17	17	17	17
[2.5,3)	52	69	52	69
[3,3.5)	27	96	27	96
[3.5,4)	4	100	4	100
[4,4.5)	0	100	0	100

In der Gruppe der Nichtraucherinnen trat ein Geburtsgewicht kleiner 2,5kg in 2,27% der Fälle auf. Bei den Raucherinnen waren es 17%.

e) Berechnen Sie die Prävalenz von Neugeborenen mit niedrigem Gewicht für Mütter, die vor der Schwangerschaft geraucht haben, und den Nichtraucherinnen.

```
gruppe1 <- neonates$gewiKat[neonates$smoke.before=="No"]
gruppe2 <- neonates$gewiKat[neonates$smoke.before=="Yes"]

# Nichtraucherinnen
jgsbook::freqTable(gruppe1)
```

Wert	Haeufig	Hkum	Relativ	Rkum
[2,2.5)	2	2	1.08	1.08
[2.5,3)	60	62	32.43	33.51
[3,3.5)	105	167	56.76	90.27
[3.5,4)	18	185	9.73	100.00
[4,4.5)	0	185	0.00	100.00

```
# Raucherinnen
jgsbook::freqTable(gruppe2)
```

Wert	Haeufig	Hkum	Relativ	Rkum
[2,2.5)	20	20	14.81	14.81
[2.5,3)	67	87	49.63	64.44
[3,3.5)	41	128	30.37	94.81
[3.5,4)	6	134	4.44	99.25
[4,4.5)	1	135	0.74	99.99

Die Prävalenz beträgt unter den Nichtraucherinnen 1,08% und unter den Raucherinnen 14,81%.

f) Berechnen Sie das relative Risiko eines niedrigen Geburtsgewichts des Neugeborenen, wenn die Mutter während der Schwangerschaft raucht, im Vergleich dazu, wenn die Mutter nicht raucht.

```
# neue binäre Variable, ob Gewicht niedrig ist
neonates$gewiLow <- FALSE
neonates$gewiLow[neonates$gewiKat=="[2,2.5)"] <- TRUE

# Kreuztabelle
table(neonates$smoke, neonates$gewiLow)
```

```
      FALSE TRUE
  No    215    5
  Yes    83   17
```

relatives Risiko $= \frac{a \cdot (c+d)}{c \cdot (a+b)}$

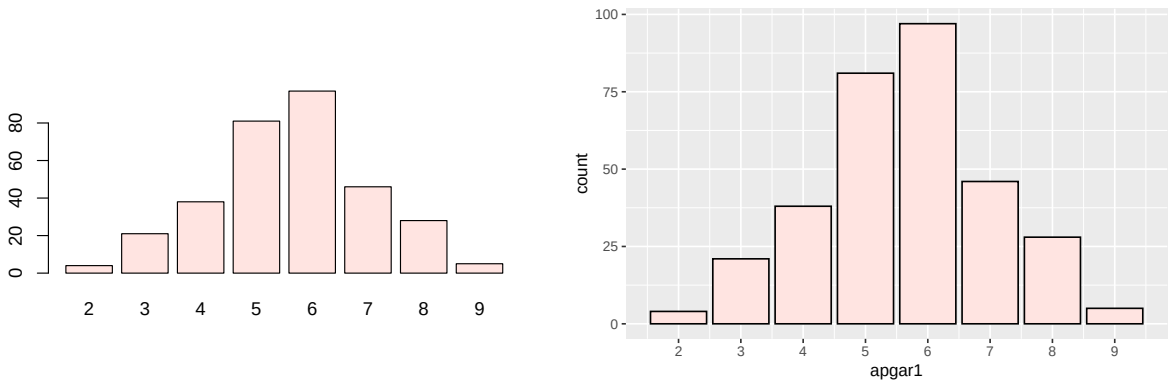
```
# Kreuztabelle als numerische Werte
tab <- as.numeric(table(neonates$smoke, neonates$gewiLow))

# rechne das relative Risiko nach der obigen Formel
( tab[1] * (tab[2]+tab[4]) ) / ( tab[2] * (tab[1]+tab[3]) )
```

```
[1] 1.177437
```

Raucherinnen haben ein 1,177437-fach höheres Risiko ein Kind mit niedrigem Gewicht zugebären als Nichtraucherinnen. Die Wahrscheinlichkeit ist in der Raucherinnengruppe also 17,74% höher als bei den Nichtraucherinnen.

g) Erstellen Sie ein Balkendiagramm des APGAR-Scores nach 1 Minute. Welcher Score ist am häufigsten?

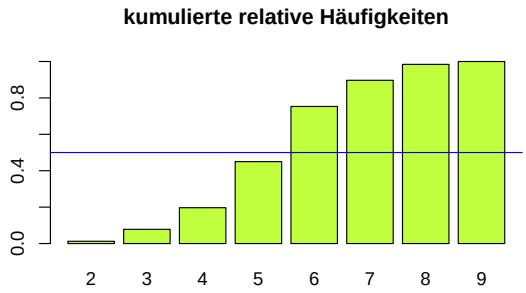
```
# mit R base
barplot(table(neonates$apgar1), col="mistyrose")
#-------------------------

# mit ggplot
ggplot(neonates, aes(x=apgar1)) +
  geom_bar(color="black", fill="mistyrose")+
  scale_x_continuous(breaks=seq(2, 9, 1))
```

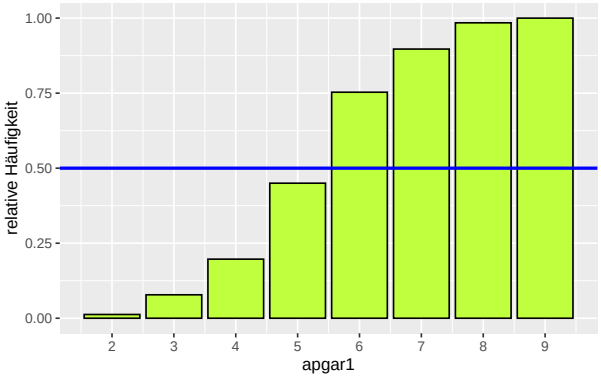


Am häufigsten tritt Wert 6 auf.

h) Erstellen Sie das Balkendiagramm der kumulierten relativen Häufigkeit des APGAR-Scores nach 1 Minute. Unter welchem Wert liegen die Hälfte der Neugeborenen?

```
# mit R base
# plotte das kumulative Balkendiagramm
barplot(cumsum(table(neonates$apgar1))/sum(table(neonates$apgar1)),
        col="olivedrab1", main = "kumulierte relative Häufigkeiten")
# Linie bei 50% ziehen
abline(h=0.5, col="blue")
#-------------------------

# mit ggplot()
ggplot(neonates, aes(x=apgar1)) +
  geom_bar(aes(y=cumsum(after_stat(count)/sum(after_stat(count)))),
        fill="olivedrab1", color="black") +
  ylab("relative Häufigkeit") +
  geom_hline(yintercept= 0.5, color="blue", linewidth=1) +
  scale_x_continuous(breaks=seq(2, 9, 1))
```

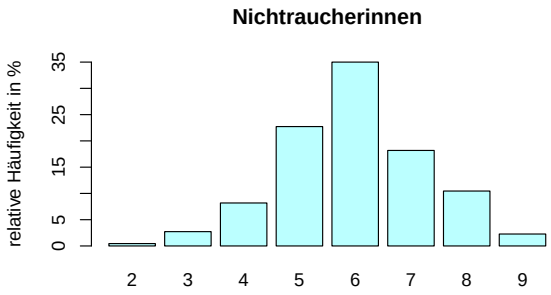


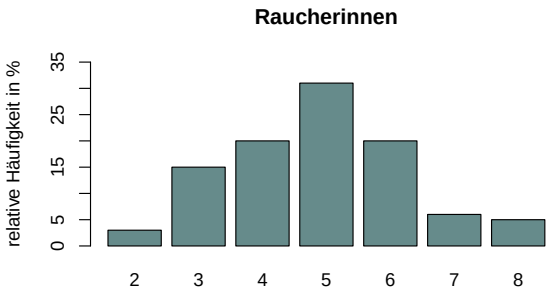


Der Median liegt bei 6.

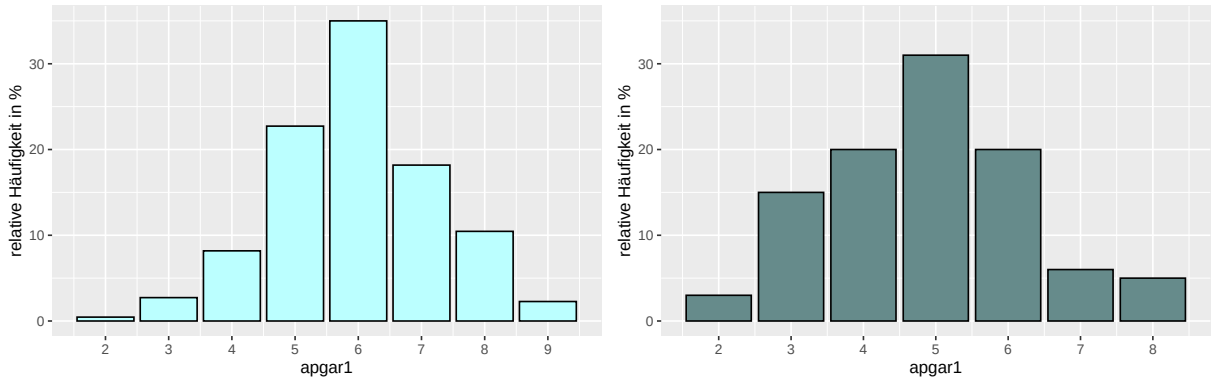
i) Vergleichen Sie die Balkendiagramme der relativen Häufigkeitsverteilungen des APGAR-Scores nach 1 Minute, je nachdem, ob die Mutter während der Schwangerschaft geraucht hat oder nicht. Welche Schlussfolgerungen können gezogen werden?

```
# mit R base
gruppe1 <- neonates$apgar1[neonates$smoke=="No"]
gruppe2 <- neonates$apgar1[neonates$smoke=="Yes"]

barplot( table(gruppe1)/sum(table(gruppe1)) *100,
        ylab="relative Häufigkeit in %", main="Nichtraucherinnen",
        ylim=c(0,35), col="paleturquoise1")
barplot( table(gruppe2)/sum(table(gruppe2)) *100,
        ylab="relative Häufigkeit in %", main="Raucherinnen",
        ylim=c(0,35), col="paleturquoise4")
```





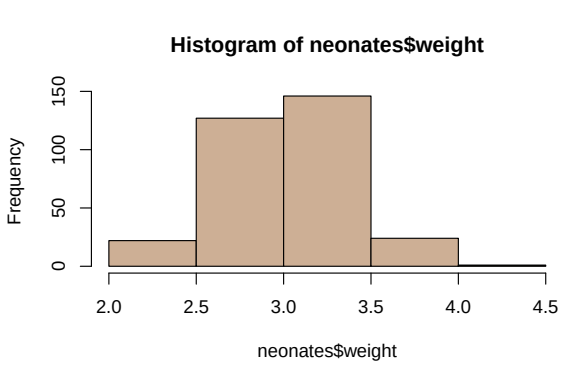
```
# mit ggplot
# Nichtraucherinnen
neonates %>%
  filter(smoke=="No") %>%
  ggplot(aes(x=apgar1))+
  geom_bar(aes(y=after_stat(count)/sum(after_stat(count))*100),
           color="black", fill="paleturquoise1")+
  scale_x_continuous(breaks=seq(2, 9, 1)) +
  ylab("relative Häufigkeit in %") +
  ylim(0,35)
#------------------------------

# Raucherinnen
neonates %>%
  filter(smoke=="Yes") %>%
  ggplot(aes(x=apgar1))+
  geom_bar(aes(y=after_stat(count)/sum(after_stat(count))*100),
           color="black", fill="paleturquoise4") +
  scale_x_continuous(breaks=seq(2, 9, 1)) +
  ylab("relative Häufigkeit in %") +
  ylim(0,35)
```

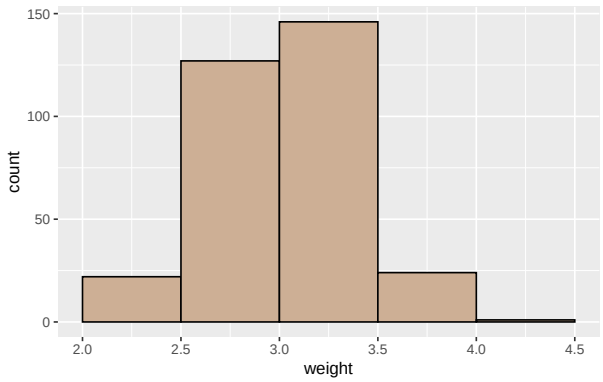


Die Kinder der Raucherinnen haben geringere APGAR-Werte.

j) Erstellen Sie ein Histogramm der Geburtsgewichte der Neugeborenen mit Klassenbreiten von 0,5 kg von 2 bis 4,5 kg. Welche Klasse enthält die meisten Neugeborenen?

```
# mit R base
hist(neonates$weight, breaks = seq(2, 4.5, 0.5),
    col="peachpuff3")
#---------------------

# mit ggplot
ggplot(neonates, aes(x=weight)) +
  geom_histogram( breaks = seq(2, 4.5, 0.5),
                  fill="peachpuff3", color="black")
```

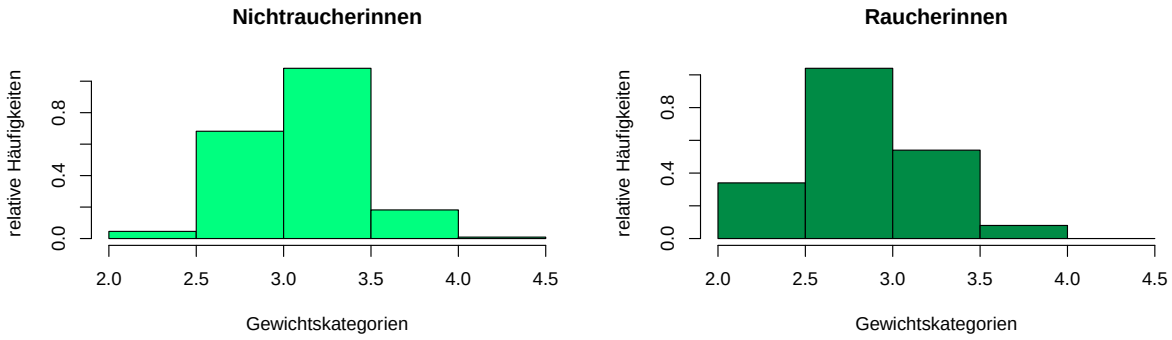




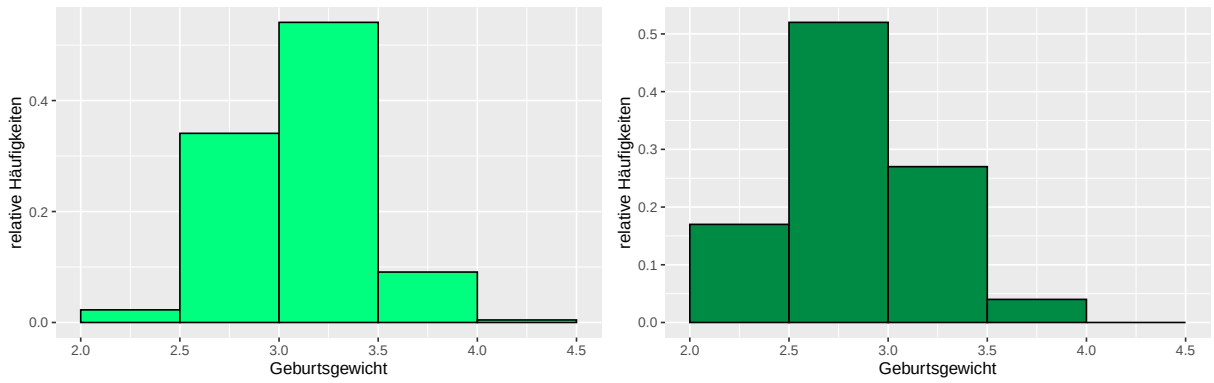
Die Gewichtsklasse 3kg-3,5kg enthält die meisten Neugeborenen

k) Vergleichen Sie die relativen Häufigkeitshistogramme der Geburtsgewichte der Neugeborenen, mit Klassenbreiten von 0,5 kg von 2 bis 4,5 kg, je nachdem, ob die Mutter während der Schwangerschaft geraucht hat oder nicht. Welche Gruppe hat Neugeborene mit geringeren Gewichten?

```
# mit R base
hist(neonates$weight[neonates$smoke=="No"],
     breaks=seq(2, 4.5, 0.5), col="springgreen1",
     main="Nichtraucherinnen", xlab="Gewichtskategorien",
     ylab="relative Häufigkeiten", freq=FALSE)
#---------------------

# Raucherinnen
hist(neonates$weight[neonates$smoke=="Yes"],
     breaks=seq(2, 4.5, 0.5), col="springgreen4",
     main="Raucherinnen", xlab="Gewichtskategorien",
     ylab="relative Häufigkeiten", freq=FALSE)
```



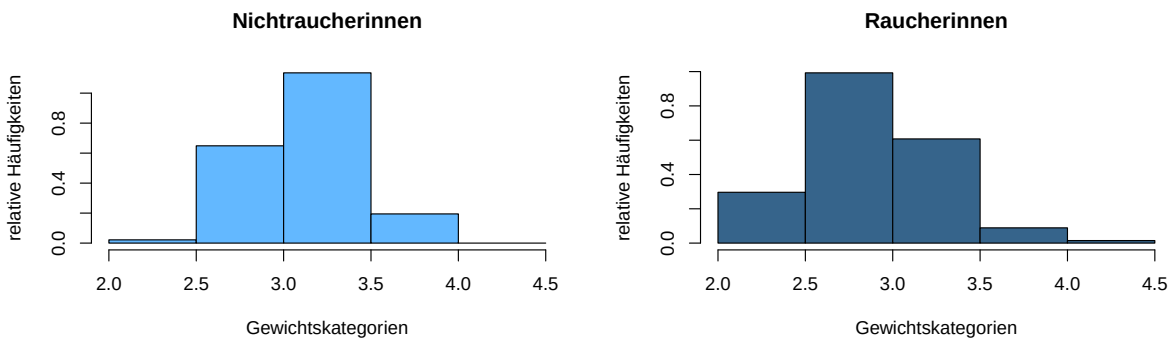

```
# mit ggplot

neonates %>%
  filter(smoke=="No") %>%
  ggplot(aes(x=weight)) +
    geom_histogram(aes(y=after_stat(count)/sum(after_stat(count))),
                   breaks=seq(2, 4.5, 0.5),
                   fill="springgreen1", color="black") +
    ylab("relative Häufigkeiten") + xlab("Geburtsgewicht")
#-----------

# Raucherinnen
neonates %>%
  filter(smoke=="Yes") %>%
  ggplot(aes(x=weight)) +
    geom_histogram(aes(y=after_stat(count)/sum(after_stat(count))),
                   breaks=seq(2, 4.5, 0.5),
                   fill="springgreen4", color="black") +
    ylab("relative Häufigkeiten") + xlab("Geburtsgewicht")
```



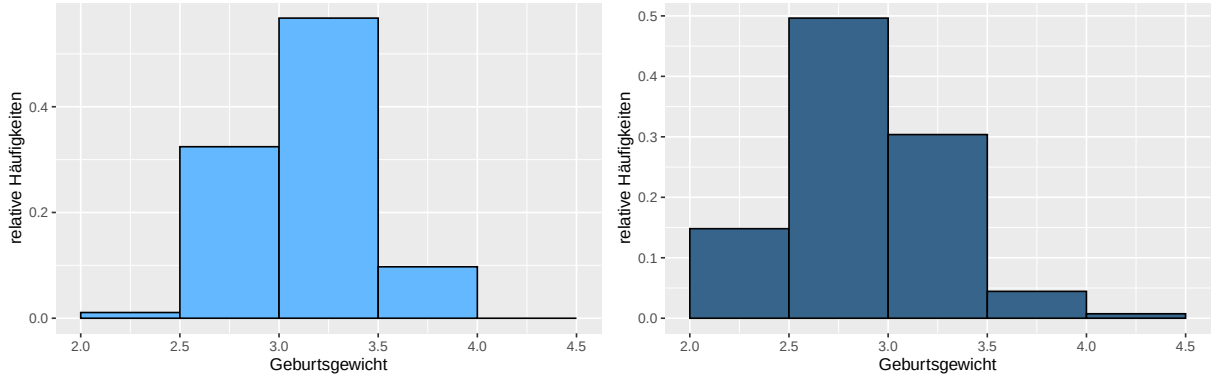
Kinder von Raucherinnen haben durchschnittlich weniger Geburtsgewicht.

l) Vergleichen Sie die relativen Häufigkeitshistogramme der Geburtsgewichte der Neugeborenen, mit Klassenbreiten von 0,5 kg von 2 bis 4,5 kg, je nachdem, ob die Mutter vor der Schwangerschaft geraucht hat oder nicht. Welche Schlussfolgerungen können gezogen werden?

```
# mit R base
hist(neonates$weight[neonates$smoke.before=="No"],
     breaks=seq(2, 4.5, 0.5), col="steelblue1",
     main="Nichtraucherinnen", xlab="Gewichtskategorien",
     ylab="relative Häufigkeiten", freq=FALSE)
#---------------------

# Raucherinnen
hist(neonates$weight[neonates$smoke.before=="Yes"],
     breaks=seq(2, 4.5, 0.5), col="steelblue4",
     main="Raucherinnen", xlab="Gewichtskategorien",
     ylab="relative Häufigkeiten", freq=FALSE)
```



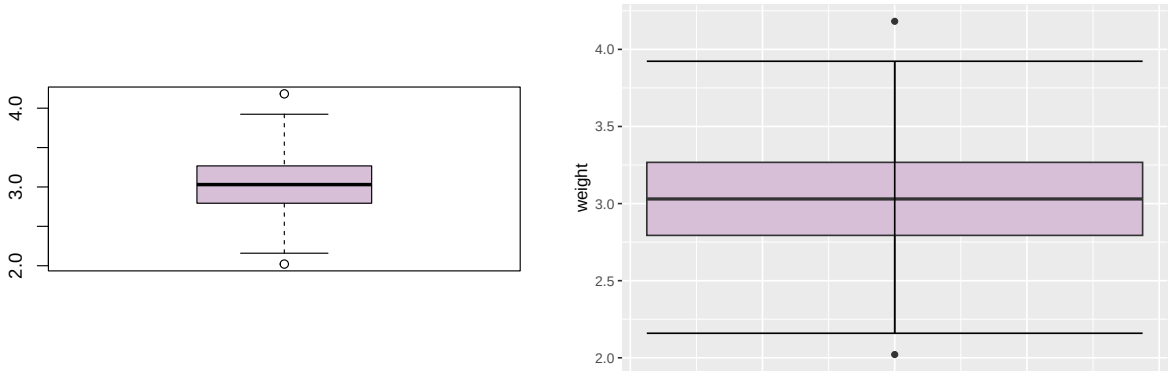

```
# mit ggplot

neonates %>%
  filter(smoke.before=="No") %>%
  ggplot(aes(x=weight)) +
    geom_histogram(aes(y=after_stat(count)/sum(after_stat(count))),
                   breaks=seq(2, 4.5, 0.5),
                   fill="steelblue1", color="black") +
    ylab("relative Häufigkeiten") + xlab("Geburtsgewicht")
#-----------

# Raucherinnen
neonates %>%
  filter(smoke.before=="Yes") %>%
  ggplot(aes(x=weight)) +
    geom_histogram(aes(y=after_stat(count)/sum(after_stat(count))),
                   breaks=seq(2, 4.5, 0.5),
                   fill="steelblue4", color="black") +
    ylab("relative Häufigkeiten") + xlab("Geburtsgewicht")
```



Kinder von Müttern, die vor der Schwangerschaft geraucht haben, haben durchschnittlich weniger Geburtsgewicht.

m) Erstellen Sie ein Boxplot der Geburtsgewichte der Neugeborenen. Welcher Gewichtsbereich kann in der Stichprobe als normal angesehen werden? Gibt es Ausreißer in der Stichprobe?

```
# mit R base
boxplot(neonates$weight, col="thistle")
# mit ggplot()
ggplot(neonates, aes(y=weight)) +
  geom_boxplot(fill="thistle") +
  stat_boxplot(geom="errorbar") +
  theme(axis.ticks.x=element_blank(),
        axis.text.x=element_blank())
```



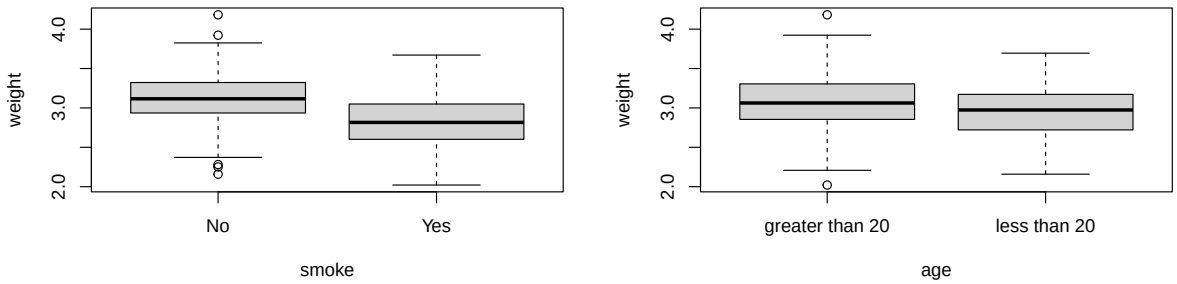

```
# Zusammenfassung
summary(neonates$weight)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.021   2.794   3.030   3.026   3.267   4.182
```

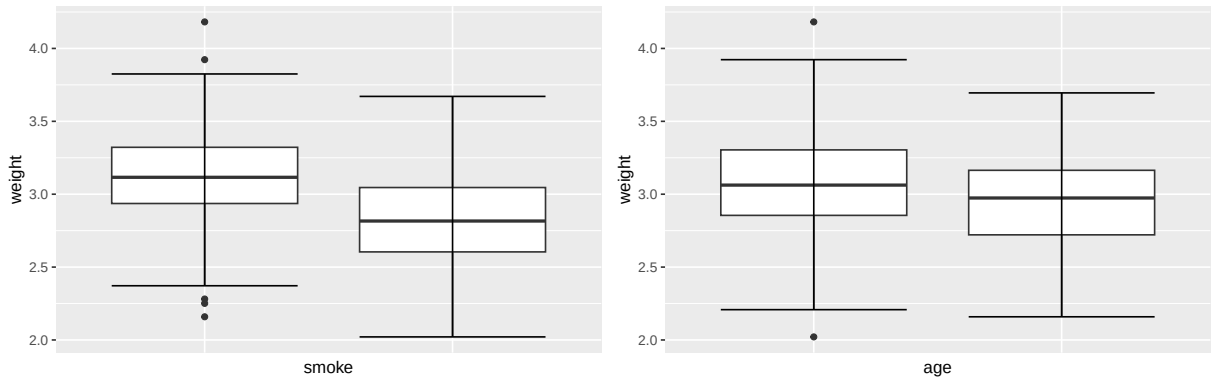
Gewichte zwischen 2,794kg und 3,267kg können als normal angesehen werden. Es gibt je einen Ausreißer nach oben und nach unten.

n) Vergleichen Sie die Boxplots der Geburtsgewichte der Neugeborenen je nachdem, ob die Mutter während der Schwangerschaft geraucht hat oder nicht und ob die Mutter unter 20 oder über 20 Jahre alt war. Welche Gruppe hat eine größere zentrale Streuung? Welche Gruppe hat Neugeborene mit geringerem Gewicht?

```
# R base
boxplot(weight ~ smoke, data=neonates)
# Alterkategorie
boxplot(weight ~ age, data=neonates)
```



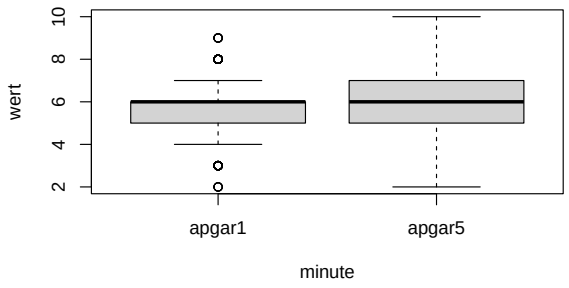

```
# ggplot Rauchen
ggplot(neonates, aes(y=weight, x=smoke)) +
  geom_boxplot() +
  stat_boxplot(geom="errorbar") +
  theme(axis.ticks.x=element_blank(),
        axis.text.x=element_blank())
#-----------------------------------
# 20 Jahre alt
ggplot(neonates, aes(y=weight, x=age)) +
  geom_boxplot() +
  stat_boxplot(geom="errorbar") +
  theme(axis.ticks.x=element_blank(),
        axis.text.x=element_blank())
```

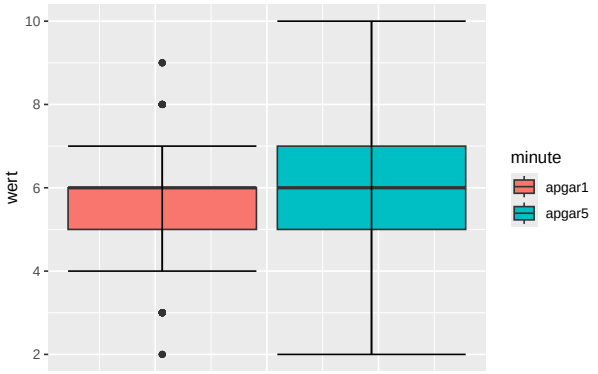


Das Gewicht ist in der Gruppe der Raucherinnen und in der Gruppe der unter-20jährigen geringer.

o) Vergleichen Sie die Boxplots der APGAR-Scores nach 1 Minute und nach 5 Minuten. Welche Variable hat eine größere zentrale Streuung?

```
# daten tidy machen
df <- pivot_longer(neonates, apgar1:apgar5,
                   names_to  = "minute",
                   values_to = "wert")

# boxplot
boxplot(wert ~ minute, data=df)
#-------------------------------

# ggplot
ggplot(df, aes(y=wert, fill=minute)) +
  geom_boxplot() +
  stat_boxplot(geom="errorbar") +
  theme(axis.ticks.x=element_blank(),
        axis.text.x=element_blank())
```





Die Streuung in `apgar5` ist größer.

2.8 Lösung zur Aufgabe 1.2.1

a) Erstellen Sie ein Datenframe mit der Variable `Kinder` und übertragen Sie die Daten.

```
# erzeuge Datenframe
df <- data.frame(Kinder = c(1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0,
                            2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2))
```

b) Berechnen Sie das arithmetische Mittel, die Varianz sowie die Standardabweichung für die Anzahl an Kindern.

```
# arithmetisches Mittel
mean(df$Kinder)
```

```
[1] 1.76
```

```
# Varianz
var(df$Kinder)
```

```
[1] 0.7733333
```

```
# Standardabweichung
sd(df$Kinder)
```

```
[1] 0.8793937
```

c) Berechnen Sie die Quartile, die Spannweite, den Interquartilsabstand, das dritte Dezil sowie das 68te Perzentil.

```
# Quartile (so wie SPSS rechnet)
quantile(df$Kinder, type=6)
```

```
  0%  25%  50%  75% 100% 
   0    1    2    2    4 
```

```
# Spannweite
range(df$Kinder)
```

```
[1] 0 4
```

```
# Interquartilsabstand (so wie SPSS rechnet)
IQR(df$Kinder, type=6)
```

```
[1] 1
```

```
# drittes Dezil und 68tes Perzentil
quantile(df$Kinder, probs=c(0.3, 0.68), type=6)
```

```
30% 68% 
  1   2 
```

2.9 Lösung zur Aufgabe 1.2.2

a) Erstellen Sie ein Datenframe mit der Variable `Patienten` und übertragen Sie die Daten.

```
# erzeuge Datenframe
df <- data.frame(Patienten = c(15, 23, 12, 10, 28, 50, 12, 17, 20,
                               21, 18, 13, 11, 12, 26, 30,  6, 16,
                               19, 22, 14, 17, 21, 28,  9, 16, 13,
                               11, 16, 20))
```

b) Berechnen Sie das arithmetische Mittel, die Varianz, die Standardabweichung und den Variationskoeffizienten.

```
# arithmetisches Mittel
mean(df$Patienten)
```

```
[1] 18.2
```

```
# Varianz
var(df$Patienten)
```

```
[1] 71.82069
```

```
# Standardabweichung
sd(df$Patienten)
```

```
[1] 8.474709
```

```
# Variationskoeffizient in Prozent
(sd(df$Patienten) / mean(df$Patienten)) * 100
```

```
[1] 46.56433
```

c) Berechnen Sie die Skewness (Schiefe) und Kurtosis ("Spitzigkeit") und interpretieren Sie die Werte.

```
# Skewness
psych::skew(df$Patienten, type=2) # rechne wie SPSS
```

```
[1] 1.902838
```

```
# Kurtosis
psych::kurtosi(df$Patienten, type=2) # rechne wie SPSS
```

```
[1] 5.796082
```

Die Skewness beträgt 1.902838, was für eine rechtsschiefe Verteilung spricht. Die Kurtosis beträgt 5.796082 und ist somit größer als 3. Das bedeutet, die Verteilung hat eine schmale Spitze und fette Ränder.

2.10 Lösung zur Aufgabe 1.2.3

a) Erstellen Sie ein Datenframe mit der Variable `Bewertung` und übertragen Sie die Daten.

```
# erzeuge Datenframe
df <- data.frame(Bewertung = c("SS", "AP", "SS", "AP", "AP", "NT", "NT", "AP",
                               "SB", "SS", "SB", "SS", "AP", "AP", "NT", "AP",
                               "SS", "NT", "SS", "NT"))
```

b) Wandeln Sie die `Bewertung` in Punkte um nach dem Schema "SS" = 2,5 | "AP" = 6 | "NT"= 8 | "SB" = 9,5.

```
# Umkodieren
df$Punkte[df$Bewertung=="SS"] <- 2.5
df$Punkte[df$Bewertung=="AP"] <- 6
df$Punkte[df$Bewertung=="NT"] <- 8
df$Punkte[df$Bewertung=="SB"] <- 9.5
```

c) Bestimmen Sie den Median und den Interquartilsabstand.

```
# Median
median(df$Punkte)
```

```
[1] 6
```

```
# Interquartilsabstand, so wie SPSS ihn rechnen würde
IQR(df$Punkte, type=6)
```

```
[1] 5.5
```

2.11 Lösung zur Aufgabe 1.2.4

a) Erstellen Sie ein Datenframe mit der Variablen `Geschlecht` und `Koerpergroesse` und übertragen Sie die Daten.

```
# erzeuge Datenframe
df <- data.frame(Geschlecht = c(rep("weiblich", 14),
                                rep("männlich", 16)),
                 Koerpergroesse = c(173, 158, 174, 166, 162, 177, 165, 154, 166,
                 182, 169, 172, 170, 168, 179, 181, 172, 194, 185, 187, 198,
                 178, 188, 171, 175, 167, 186, 172, 176, 187))
```

b) Bestimmen Sie in Abhängigkeit zum `Geschlecht` das arithmetische Mittel, den Median, die Varianz, die Standardabweichung sowie die Quartile.

```
# mit dplyr
df %>% group_by(Geschlecht) %>%
  reframe(aritMittel = mean(Koerpergroesse),
          Median = median(Koerpergroesse),
          Varianz = var(Koerpergroesse),
          StdAbw  = sd(Koerpergroesse),
          Q1 = quantile(Koerpergroesse, probs=0.25, type=6),
          Q3 = quantile(Koerpergroesse, probs=0.75, type=6)
  )
```

Geschlecht	aritMittel	Median	Varianz	StdAbw	Q1	Q3
männlich	181.0000	180.0	76.80000	8.763561	172.75	187.00
weiblich	168.2857	168.5	54.37363	7.373848	164.25	173.25

2.12 Lösung zur Aufgabe 1.2.5

a) Bestimmen Sie das arithmetische Mittel, den Median, die Varianz sowie die Standardabweichung der Verletzungen.

```
# Daten übertragen
Handball <- c(0, 1, 2, 1, 3, 0, 1, 0, 1, 2, 0, 1, 1,
              1, 2, 0, 1, 3, 2, 1, 2, 1, 0, 1)

# Mittelwert
mean(Handball)
```

```
[1] 1.125
```

```r
# Median
median(Handball)
```

```
[1] 1
```

```r
# Varianz
var(Handball)
```

```
[1] 0.8097826
```

```r
# Standardabweichung
sd(Handball)
```

```
[1] 0.8998792
```

b) Bestimmen Sie die Skewness und Kortosis der Verteilung.

```r
# Skewness, wie SPSS
psych::skew(Handball, type=2)
```

```
[1] 0.5186785
```

```r
# Kurtosis, wie SPSS
psych::kurtosi(Handball, type=2)
```

```
[1] -0.226357
```

c) Berechnen Sie das vierte und achte Dezil der Verteilung.

```r
quantile(Handball, probs=c(0.4, 0.8), type=6)
```

```
40% 80%
  1   2
```

2.13 Lösung zur Aufgabe 1.2.6

Welcher Monitor funktioniert besser?

```
# Daten übertragen
monitor <- data.frame(Unterarm=c(111, 109, 112, 111, 113, 113, 114, 111),
                      Handgelenk=c(115, 113, 117, 116, 112, 112, 117, 112)
)

# Standardabweichungen vergleichen
sd(monitor$Unterarm)
```

```
[1] 1.581139
```

```
sd(monitor$Handgelenk)
```

```
[1] 2.251983
```

Die Daten des Monitors am Handgelenk haben eine größere Streuung als jene vom Monitor am Unterarm.

2.14 Lösung zur Aufgabe 1.2.7

a) Bestimmen Sie das arithmetische Mittel, den Median, die Varianz sowie die Standardabweichung des `Alters` für jeden `Familienstand`.

```
# Daten übertragen
df <- data.frame(Familienstand = c(rep("Single", 9),
                                   rep("Verheiratet", 7),
                                   rep("Verwitwet", 7),
                                   rep("Geschieden", 5)),
                 Alter = c(31, 45, 35, 65, 21, 38, 62, 22, 31,
                           72, 39, 62, 59, 25, 44, 54,
                           80, 68, 65, 40, 78, 69, 75,
                           31, 65, 59, 58, 50)
                 )

# dplyr
df %>% group_by(Familienstand) %>%
       reframe(Mittel = mean(Alter),
               Median = median(Alter),
               Varianz = var(Alter),
               StdAbw = sd(Alter))
```

Familienstand	Mittel	Median	Varianz	StdAbw
Geschieden	52.60000	58	174.3000	13.20227
Single	38.88889	35	249.8611	15.80700
Verheiratet	50.71429	54	250.5714	15.82945
Verwitwet	67.85714	69	181.1429	13.45893

Welche Gruppe hat den "besten" Mittelwert?

```
df %>% group_by(Familienstand) %>%
       reframe(Mittel = mean(Alter),
               Median = median(Alter),
               Varianz = var(Alter),
               StdAbw = sd(Alter),
               Skew = psych::skew(Alter, type=2),
               Kurto = psych::kurtosi(Alter, type=2))
```

Familienstand	Mittel	Median	Varianz	StdAbw	Skew	Kurto
Geschieden	52.60000	58	174.3000	13.20227	-1.4067277	2.0349803
Single	38.88889	35	249.8611	15.80700	0.7646514	-0.5243790
Verheiratet	50.71429	54	250.5714	15.82945	-0.4251562	-0.3129758
Verwitwet	67.85714	69	181.1429	13.45893	-1.7649555	3.6623934

In der Gruppe der Geschiedenen ist die Streuung am geringsten, aber das arithmetische Mittel ist weit vom Median entfernt. Insofern scheint der Mittelwert der Verwitweten am "besten" zu sein, da er sowohl eine geringe Differenz zum Median als auch eine niedrige Varianz und Standardabweichung aufweist. Die Gruppe der Verheirateten weist hingegen die geringste Schiefe auf. Ach hier könnten wir argumentieren, dass dies für den "besten" Mittelwert spräche.

2.15 Lösung zur Aufgabe 1.2.8

a) Vergleichen Sie das arithmetische Mittel, die Standardabweichung, die Skewness und Kurtosis des `Blutdrucks` zwischen Rauchern und Nichtrauchern.

```
# Daten übertragen
df <- data.frame(Rauchen = c("ja", "nein", "ja", "ja", "ja", "nein", "nein",
                             "ja", "nein", "ja", "nein", "ja", "nein",
                             "ja", "nein", "nein", "ja", "nein", "nein",
                             "nein", "ja", "nein", "ja", "nein", "ja"  ),
                 Alkohol = c("nein", "nein", "ja", "ja", "nein", "nein", "ja",
                             "ja", "nein", "ja", "nein", "ja", "ja" , "ja",
                             "nein", "ja", "ja", "nein", "nein", "ja", "ja",
                             "ja", "nein", "ja", "nein" ),
                 Blutdruck = c(80, 92, 75, 56, 89, 93, 101, 67, 89, 63, 98, 58,
                               91, 71, 52, 98, 104, 57, 89, 70, 93, 69, 82, 70,
                               49 )
                 )

# dplyr
df %>% group_by(Rauchen) %>%
       reframe(Mittel = mean(Blutdruck),
               StdAbw = sd(Blutdruck),
               Skew =  psych::skew(Blutdruck, type=2),
               Kurto = psych::kurtosi(Blutdruck, type=2))
```

Rauchen	Mittel	StdAbw	Skew	Kurto
ja	73.91667	16.43421	0.2814578	-0.6239070
nein	82.23077	16.46792	-0.6998100	-0.9345984

b) Vergleichen Sie die selben Werte zwischen der Alkohol- und Nicht-Alkoholgruppe.

```
# dplyr
df %>% group_by(Alkohol) %>%
       reframe(Mittel = mean(Blutdruck),
               StdAbw = sd(Blutdruck),
               Skew =  psych::skew(Blutdruck, type=2),
               Kurto = psych::kurtosi(Blutdruck, type=2))
```

Alkohol	Mittel	StdAbw	Skew	Kurto
ja	77.57143	16.39284	0.4549274	-1.3206125
nein	79.09091	17.74517	-0.9484563	-0.8040328

c) Vergleichen Sie die selben Werte zwischen der Raucher- und Alkoholgruppe, zwischen der Raucher- und Nicht-Alkoholgruppe, der Nichtraucher- und Alkoholgruppe sowie der Nichtraucher- und Nicht-Alkoholgruppe.

```
# dplyr
df %>% group_by(Alkohol, Rauchen) %>%
       reframe(Mittel = mean(Blutdruck),
               StdAbw = sd(Blutdruck),
               Skew =  psych::skew(Blutdruck, type=2),
               Kurto = psych::kurtosi(Blutdruck, type=2))
```

Alkohol	Rauchen	Mittel	StdAbw	Skew	Kurto
ja	ja	73.37500	16.97845	1.0124993	0.0026964
ja	nein	83.16667	15.14486	0.1733778	-2.8742732
nein	ja	75.00000	17.75763	-1.7101307	3.2083168
nein	nein	81.42857	18.69810	-1.1391312	-0.7477992

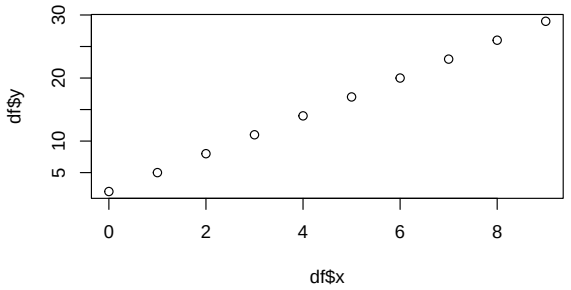
2.16 Lösung zur Aufgabe 1.3.1

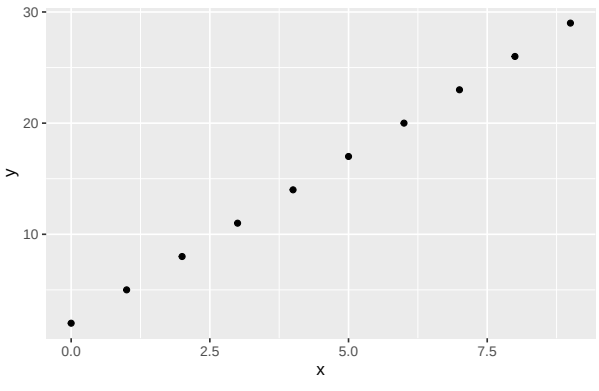
a) Erstellen Sie ein Datenframe mit den Variablen `x` und `y`.

```
# erzeuge Datenframe
df <- data.frame(x = c( 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 ),
                 y = c( 2, 5, 8, 11, 14, 17, 20, 23, 26, 29))
```

b) Erzeugen Sie ein Scatterplot von `x` und `y`. Bestimmen Sie anhand des Plots, welche Regressionsfunktion die Daten am besten erklären würde.

```
# plot()
plot(df$x, df$y)
# ggplot()
ggplot(df, aes(x=x, y=y)) +
  geom_point()
```





Es ist ein deutlicher linearer Zusammenhang erkennbar.

c) Führen Sie die Regression durch.

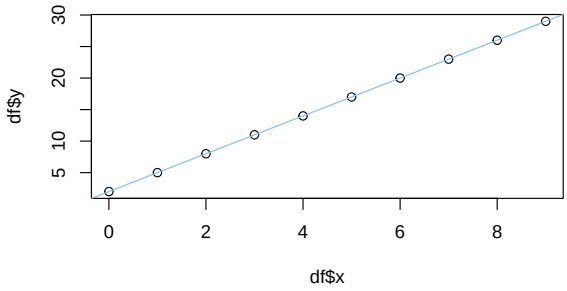
```
# lineares Modell
fit <- lm(y ~ x, data=df)

# anschauen
summary(fit)
```

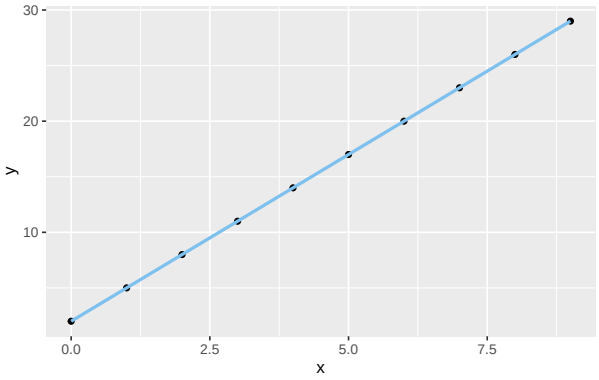
```
Warning in summary.lm(fit): im Wesentlichen ein perfekter Fit: summary kann
unzuverlässig sein

Call:
lm(formula = y ~ x, data = df)

Residuals:
       Min         1Q     Median         3Q        Max
-3.675e-15 -8.783e-16  5.168e-16  9.646e-16  1.944e-15

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)
(Intercept) 2.000e+00  1.049e-15 1.906e+15   <2e-16 ***
x           3.000e+00  1.965e-16 1.527e+16   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.785e-15 on 8 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1
F-statistic: 2.33e+32 on 1 and 8 DF,  p-value: < 2.2e-16
```

d) Fügen Sie die Regressionsfunktion `y erklärt durch x` dem Plot hinzu.

```
# plot()
plot(df$x, df$y)
# Regressionsgerade
abline(lm(y~x, data=df), col="skyblue2")
# ggplot()
ggplot(df, aes(x=x, y=y)) +
  geom_point() +
  geom_smooth(method="lm", color="skyblue2")
```

```
`geom_smooth()` using formula = 'y ~ x'
```

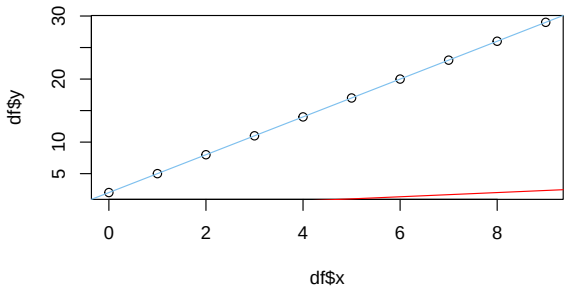


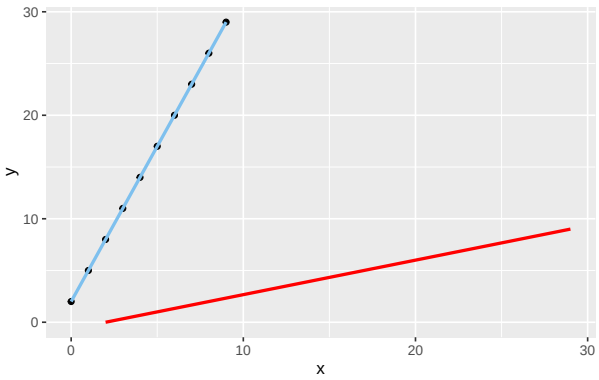


e) Fügen Sie die Regressionsfunktion `x erklärt durch y` ebenfalls dem Plot hinzu, aber in roter Farbe.

```
# plot()
plot(df$x, df$y)
# Regressionsgeraden
abline(lm(y~x, data=df), col="skyblue2")
abline(lm(x~y, data=df), col="red")
# ggplot()
ggplot(df, aes(x=x, y=y)) +
  geom_point() +
  geom_smooth(method="lm", color="skyblue2") +
  geom_smooth(aes(x=y, y=x), method="lm", color="red")
```

```
`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
```





Achten Sie darauf, die Variablen nicht zu vertauschen!

f) Wie große sind die Residuen?

```
fit <- lm(y ~ x, data=df)

# Residuen
fit$residuals
```

```
            1             2             3             4             5
-3.675227e-15  4.362024e-16  1.944285e-15  1.385502e-15  1.048764e-15
            6             7             8             9            10
 7.120252e-16  5.973314e-16 -1.293719e-15  3.679437e-16 -1.523107e-15
```

Die Residuen sind sehr klein. Die Regressionsgerade scheint sehr gut zu passen.

2.17 Lösung zur Aufgabe 1.3.2

a) Erstellen Sie ein Datenframe mit den Variablen `Lernen` und `Durchgefallen`.

```
# erzeuge Datenframe
df <- data.frame(Lernen = c( 3.5, 0.6, 2.8, 2.5, 2.6, 3.9, 1.5, 0.7, 3.6, 3.7,
                             2.2, 3.3, 1.7, 1.1, 2.0, 3.5, 2.1, 1.8, 1.1, 0.7,
                             1.3, 3.1, 2.3, 3.2, 0.9, 1.7, 0.2, 2.9, 1.0, 2.3),
                 Durchgefallen = c( 1, 5, 1, 3, 1, 0, 3, 3, 1, 1,
                                    2, 0, 3, 3, 3, 0, 2, 2, 4, 4,
                                    4, 0, 2, 2, 4, 2, 5, 1, 3, 2))
```

b) Erzeugen Sie eine Kreuztabelle der Variablen `Lernen` und `Durchgefallen`.

```
# entweder
table(df$Lernen, df$Durchgefallen)
```

```
     0 1 2 3 4 5
  0.2 0 0 0 0 0 1
  0.6 0 0 0 0 0 1
  0.7 0 0 0 1 1 0
  0.9 0 0 0 0 1 0
  1   0 0 0 1 0 0
  1.1 0 0 0 1 1 0
  1.3 0 0 0 0 1 0
  1.5 0 0 0 1 0 0
  1.7 0 0 1 1 0 0
  1.8 0 0 1 0 0 0
  2   0 0 0 1 0 0
  2.1 0 0 1 0 0 0
  2.2 0 0 1 0 0 0
  2.3 0 0 2 0 0 0
```

```
  2.5 0 0 0 1 0 0
  2.6 0 1 0 0 0 0
  2.8 0 1 0 0 0 0
  2.9 0 1 0 0 0 0
  3.1 1 0 0 0 0 0
  3.2 0 0 1 0 0 0
  3.3 1 0 0 0 0 0
  3.5 1 1 0 0 0 0
  3.6 0 1 0 0 0 0
  3.7 0 1 0 0 0 0
  3.9 1 0 0 0 0 0
```

```
# oder
xtabs(~ Lernen + Durchgefallen, data=df)
```

```
      Durchgefallen
Lernen 0 1 2 3 4 5
   0.2 0 0 0 0 0 1
   0.6 0 0 0 0 0 1
   0.7 0 0 0 1 1 0
   0.9 0 0 0 0 1 0
   1   0 0 0 1 0 0
   1.1 0 0 0 1 1 0
   1.3 0 0 0 0 1 0
   1.5 0 0 0 1 0 0
   1.7 0 0 1 1 0 0
   1.8 0 0 1 0 0 0
   2   0 0 0 1 0 0
   2.1 0 0 1 0 0 0
   2.2 0 0 1 0 0 0
   2.3 0 0 2 0 0 0
   2.5 0 0 0 1 0 0
   2.6 0 1 0 0 0 0
   2.8 0 1 0 0 0 0
   2.9 0 1 0 0 0 0
   3.1 1 0 0 0 0 0
   3.2 0 0 1 0 0 0
   3.3 1 0 0 0 0 0
   3.5 1 1 0 0 0 0
   3.6 0 1 0 0 0 0
   3.7 0 1 0 0 0 0
   3.9 1 0 0 0 0 0
```

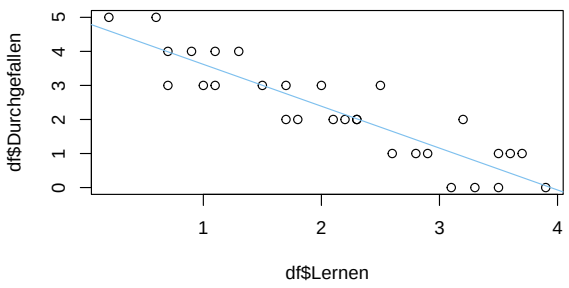
c) Führen Sie eine lineare Regression `Durchgefallen erklärt durch Lernen` durch und plotten Sie Ihr Ergebnis.

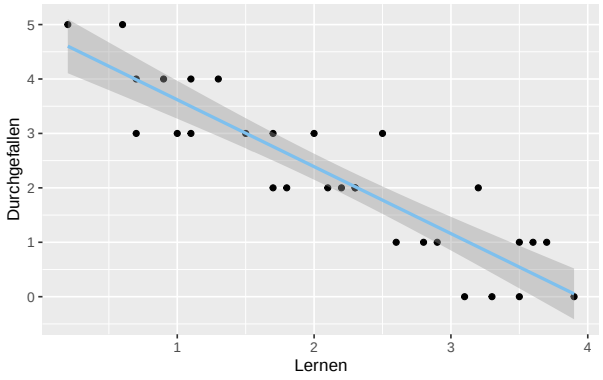
```
# lineare Regression
fit <- lm(Durchgefallen ~ Lernen , data=df)
summary(fit)
```

```
Call:
lm(formula = Durchgefallen ~ Lernen, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.03614 -0.53214 -0.02013  0.49187  1.22587

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.8491     0.2622   18.49  < 2e-16 ***
Lernen       -1.2300     0.1106  -11.12  8.7e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6359 on 28 degrees of freedom
Multiple R-squared:  0.8155,    Adjusted R-squared:  0.8089
F-statistic: 123.8 on 1 and 28 DF,  p-value: 8.7e-12
```

```
# plot()
plot(df$Lernen, df$Durchgefallen)
# Regressionsgerade
abline(lm(Durchgefallen ~ Lernen, data=df), col="skyblue2")
# ggplot()
ggplot(df, aes(x=Lernen, y=Durchgefallen)) +
  geom_point() +
  geom_smooth(method="lm", color="skyblue2")
```

```
`geom_smooth()` using formula = 'y ~ x'
```





d) Wie lauten die Regressionskoeffizient des Modells, und wie ist er zu interpretieren?

```
# Koeffizienten anzeigen
fit$coefficients
```

```
(Intercept)      Lernen
   4.849127   -1.229997
```

Der Regressionskoeffizient für `Lernen` beträgt -1.2299972. Das bedeutet, dass mit ungefähr jeder Stunde Lernen ein Kurs weniger nicht bestanden wird.

e) Ist das soeben erstellte Modell *besser* als das in Abschnitt 2.16 berechnete? Vergleichen Sie zur Beantwortung die Residuen beider Modelle.

```
# aktuelles Modell
# Residuen anschauen
fit$residuals
```

```
          1           2           3           4           5           6
 0.455862792  0.888870974 -0.405135233  1.225865613 -0.651134669 -0.052138337
          7           8           9          10          11          12
-0.004131565 -0.988129308  0.578862510  0.701862227 -0.143133540 -0.790136644
         13          14          15          16          17          18
 0.241867871 -0.496130437  0.610867024 -0.544137208 -0.266133258 -0.635132412
         19          20          21          22          23          24
 0.503869563  0.011870692  0.749868999 -1.036136080 -0.020133822  1.086863638
         25          26          27          28          29          30
 0.257870128 -0.758132129  0.396872103 -0.282135515 -0.619130154 -0.020133822
```

```
# Modell aus anderer Aufgabe
df2 <- data.frame(x = c( 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 ),
                  y = c( 2, 5, 8, 11, 14, 17, 20, 23, 26, 29))

fit2 <- lm(y~x, data=df2)
fit2$residuals
```

```
           1            2            3            4            5
-3.675227e-15  4.362024e-16  1.944285e-15  1.385502e-15  1.048764e-15
           6            7            8            9           10
 7.120252e-16  5.973314e-16 -1.293719e-15  3.679437e-16 -1.523107e-15
```

Im aktuellen Modell sind die Residuen größer als im vorherigen Modell. Somit ist das vorherige Modell besser.

f) Berechnen Sie den linearen Bestimmungskoeffizient und den Korrelationskoeffizient. Ist das lineare Modell ein gutes Modell, um die Beziehung zwischen den gescheiterten Prüfungen und den täglichen Studienzeiten zu erklären? Wie viel Prozent der Variabilität der durchgefallenen Prüfungen wird durch das lineare Modell erklärt?

```
# aktuelles Modell
lernen <- summary(fit)
# R^2 anschauen
lernen$r.squared
```

```
[1] 0.8154995
```

```
# Korrelationskoeffizient
cor.test(df$Lernen, df$Durchgefallen)
```

```
    Pearson's product-moment correlation

data:  df$Lernen and df$Durchgefallen
t = -11.125, df = 28, p-value = 8.7e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9532031 -0.8045264
sample estimates:
       cor
-0.9030501
```

Das Bestimmtheitsmaß R^2 beträgt 0.8154995. Somit können 81.55% des Rauschens im aktuellen Modell erklärt werden.
Der Korrelationskoeffizient von -0.9030501 ist nahe an -1. Dies spricht für einen starken negativen Zusammenhang.

g) Benutzen Sie das lineare Modell, um die Anzahl an durchgefallenen Prüfungen für einen Studenten zu bestimmen, der 3 Stunden Lernzeit investiert hat. Wie glaubwürdig ist die Vorhersage?

```
# aktuelles Modell
predict(fit, list(Lernen=3))
```

```
       1
1.159136
```

Wenn der Student 3 Stunden lernt, wird er wahrscheinlich “nur” durch 1 Kurs durchfallen.

h) Wie viele Stunden Lernzeit wird benötigt, um alle Kurse zu bestehen?

```
# neues Modell
fit <- lm(Lernen ~ Durchgefallen, data = df)
# Wieviel lernen für Durchgefallen=0?
predict(fit, list(Durchgefallen=0))
```

```
       1
3.607387
```

Wenn der Student 3 Stunden lernt, wird er wahrscheinlich "nur" durch 1 Kurs durchfallen.

2.18 Lösung zur Aufgabe 1.3.3

a) Erstellen Sie ein Datenframe mit den Variablen `Minuten` und `Alkohol`.

```
# erzeuge Datenframe
df <- data.frame(Alkohol = c(1.6, 1.7, 1.5, 1.1, 0.7, 0.2, 2.1),
                 Minuten = c(30, 60, 90, 120, 150, 180, 210))
```

b) Bestimmen Sie den passenden Korrelationskoeffizienten. Werden die Daten ausreichend gut durch das Modell beschrieben?

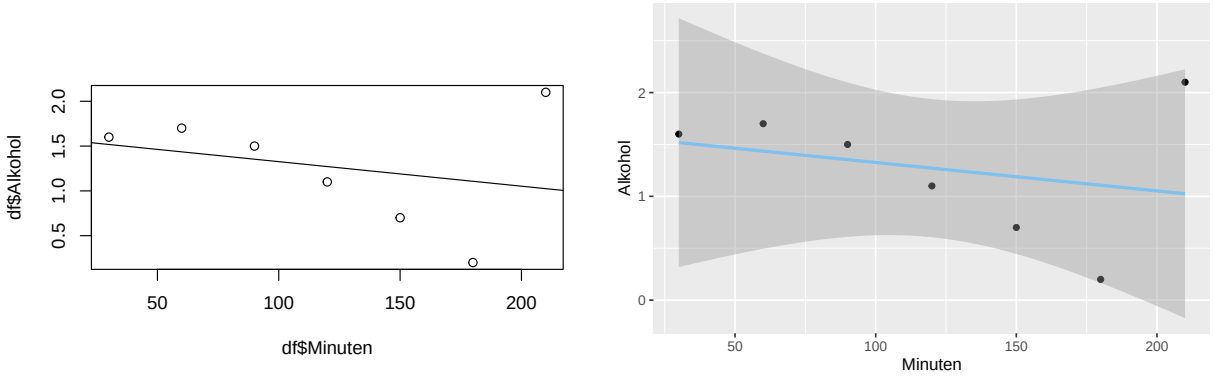
```
# Korrelation
cor(df$Minuten, df$Alkohol)
```

```
[1] -0.2730367
```

Der Korrelationskoeffizient ist eher gering. Das spricht für keinen starken Zusammenhang.

c) Plotten Sie das lineare Regressionsmodell `Alkohol erklärt durch Minuten`. Gibt es Punkte mit großen Residuen? Wenn ja, entfernen Sie diese und führen die Berechnungen erneut durch. Hat sich der Korrelationskoeffizient verbessert?

```
# plot()
plot(df$Minuten, df$Alkohol)
abline(lm(Alkohol ~ Minuten, data=df))
# ggplot()
ggplot(df, aes(x=Minuten, y=Alkohol)) +
  geom_point()+
  geom_smooth(method="lm", color="skyblue2")
```



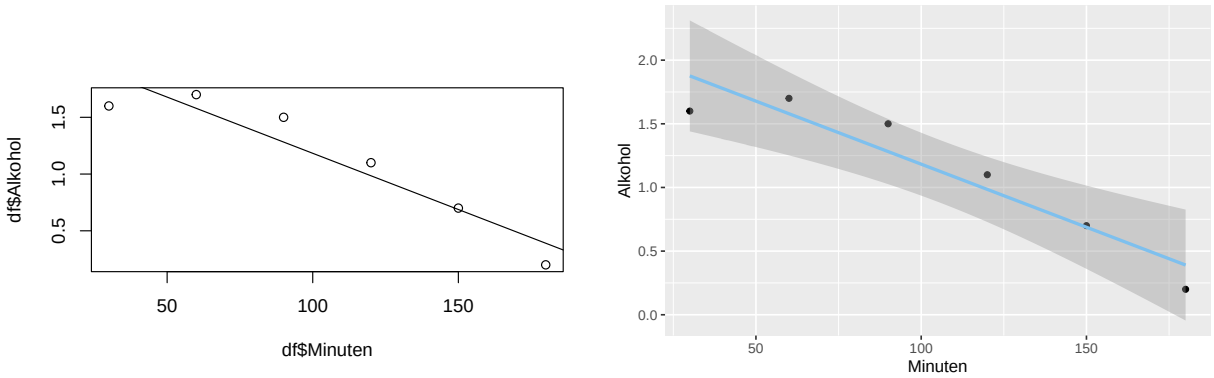
Der letzte Wert ist ein deutlicher Ausreißer, wahrscheinlich ein Tippfehler bei der Dateneingabe.

```
# entferne letzten Wert
df <- df[-7,]
# Korrelation
cor(df$Minuten, df$Alkohol)
```

```
[1] -0.944155
```

Der Korrelationskoeffizient ist nun sehr nah an -1. Das spricht für einen starken Zusammenhang.

```
# Modell
# plot()
plot(df$Minuten, df$Alkohol)
abline(lm(Alkohol ~ Minuten, data=df))
# ggplot()
ggplot(df, aes(x=Minuten, y=Alkohol)) +
  geom_point()+
  geom_smooth(method="lm", color="skyblue2")
```



d) Mit welcher Geschwindigkeit wird der Alkohol pro Minute verstoffwechselt?

```
# Koeffizienten
fit <- lm(Alkohol ~ Minuten, data=df)
fit$coefficient
```

```
 (Intercept)      Minuten
2.173333333 -0.009904762
```

Der Alkoholspiegel sinkt pro Minute um -0.0099048 g/l.

e) Wenn es gesetzlich erlaubt wäre, mit einem Blutalkoholwert von $0,3$ g/l Auto zu fahren, wie lange muss die Person warten, nachdem sie 1 Liter Weingetrunken hat, um wieder fahrtüchtig zu sein? Wie zuverlässig ist diese Vorhersage?

```
# Koeffizienten
fit <- lm(Minuten ~ Alkohol, data=df)
predict(fit, list(Alkohol=0.3))
```

```
  1
180
```

Der Alkoholspiegel wird nach 180 Minuten auf $0,3$ g/l fallen.

2.19 Lösung zur Aufgabe 1.3.4

a) Laden Sie den Datensatz `age.height` in Ihre R-Session.

```
# lade Datensatz
load(url("https://www.produnis.de/R/data/age.height.RData"))
```

b) Berechnen Sie die Regressionsgerade `Größe erklärt durch Alter`. Ist das lineare Modell geeignet, den Zusammenhang zwischen `Alter` und `Körpergröße` zu erklären?

```
# Regression
fit <- lm(height ~ age, data=age.height)
summary(fit)
```

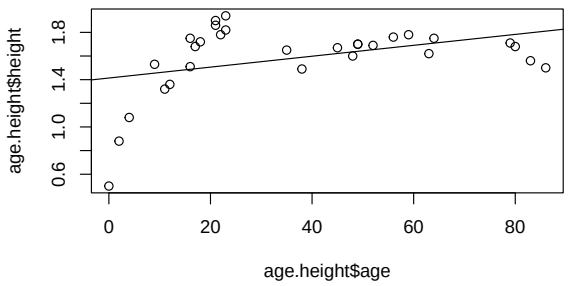
```
Call:
lm(formula = height ~ age, data = age.height)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9137 -0.1018  0.0449  0.1644  0.4202

Coefficients:
```

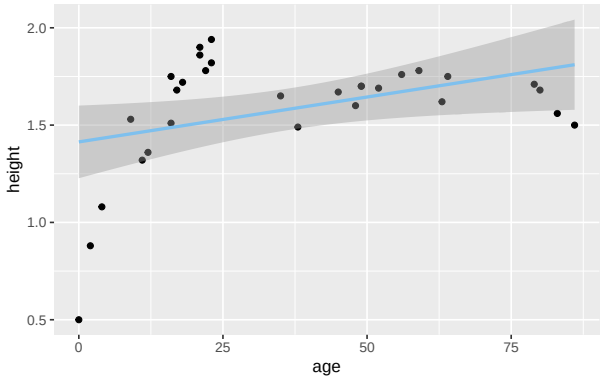
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.413724   0.091080  15.522 2.77e-15 ***
age         0.004612   0.002036   2.265   0.0314 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2852 on 28 degrees of freedom
Multiple R-squared:  0.1549,    Adjusted R-squared:  0.1247
F-statistic: 5.131 on 1 and 28 DF,  p-value: 0.03142
```

R^2 ist eher gering, es können nur 15.49% des Rauschens mit dem Modell erklärt werden.

c) Erstellen Sie eine Punktwolke inklusive der Regressionsgeraden. Ab welchem Alter ändert sich die Punktetendenz?

```
# plot()
plot(age.height$age, age.height$height)
abline(fit)
# ggplot()
ggplot(age.height, aes(x=age, y=height)) +
  geom_point()+
  geom_smooth(method="lm", color="skyblue2")
```





Ab etwa 20 Jahren ändert sich die Punktetendenz.

d) Erstellen Sie eine Gruppierungsvariable, welche `Alter` in einen ordinalen Faktor mit den Ausprägungen "`jünger als 20`" und "`20 und älter`" einteilt.

```
# klassieren
age.height$ageK <- cut(age.height$age,
                       breaks = c(0,20, Inf),
                       right=FALSE,
                       labels = c("jünger als 20",
                                  "20 und älter"))
# anschauen

head(age.height)
```

age	height	ageK
18	1.72	jünger als 20
21	1.90	20 und älter
45	1.67	20 und älter
59	1.78	20 und älter
21	1.86	20 und älter
22	1.78	20 und älter

e) Führen Sie die lineare Regressionsanalyse für beide Gruppen erneut durch. In welcher Gruppe wird der Zusammenhang zwischen `Alter` und `Körpergröße` am besten erklärt?

```
# Gruppen
df1 <- subset(age.height, ageK=="jünger als 20")
# Regression
fit1 <- lm(height ~ age, data=df1)
summary(fit1)
```

```
Call:
lm(formula = height ~ age, data = df1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22746 -0.05601 -0.03485  0.08416  0.28351

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.727459   0.094487   7.699 5.75e-05 ***
age         0.057671   0.007738   7.453 7.25e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1525 on 8 degrees of freedom
Multiple R-squared:  0.8741,    Adjusted R-squared:  0.8584
```

```
F-statistic: 55.54 on 1 and 8 DF,  p-value: 7.245e-05
```

Das Bestimmtheitsmaß in der Gruppe "jünger als 20" liegt bei 0.8741033, d.h. es werden 87.41% des Rauschens erklärt.

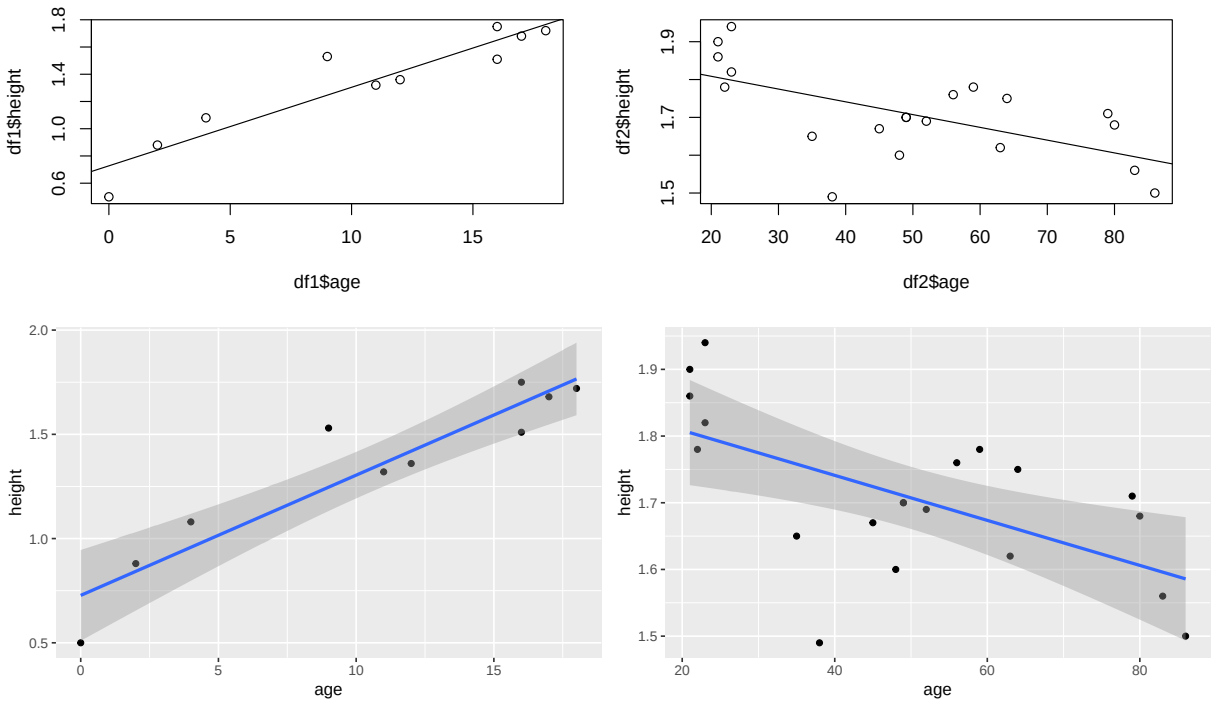
```
# Gruppen
df2 <- subset(age.height, ageK=="20 und älter")
# Regression
fit2 <- lm(height ~ age, data=df2)
summary(fit2)
```

```

Call:
lm(formula = height ~ age, data = df2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25783 -0.04614 -0.01064  0.07793  0.14155

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.876084   0.056839  33.007  < 2e-16 ***
age         -0.003375   0.001051  -3.213  0.00483 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09931 on 18 degrees of freedom
Multiple R-squared:  0.3644,    Adjusted R-squared:  0.3291
F-statistic: 10.32 on 1 and 18 DF,  p-value: 0.004827
```

Das Bestimmtheitsmaß in der Gruppe "20 und älter" liegt bei 0.3644171, d.h. es werden 36.44% des Rauschens erklärt.

f) Plotten Sie Ihre Modelle

```
# < 20 Jahre
plot(df1$age, df1$height)
abline(fit1)
# >= 20 Jahre
plot(df2$age, df2$height)
abline(fit2)
## ggplot()
# < 20 Jahre
ggplot(df1, aes(x=age, y=height)) +
  geom_point() +
  geom_smooth(method="lm")
# >= 20 Jahre
ggplot(df2, aes(x=age, y=height)) +
  geom_point() +
  geom_smooth(method="lm")
```



g) Welche Körpergröße sagt Ihr Modell für eine 14jährige Person vorher, und welche für eine 38jährige Person?

```
# 14 jährige Person
predict(fit1, list(age=14))
```

```
       1
```

```
1.534847
```

```
# 38 jährige Person
predict(fit2, list(age=38))
```

```
       1
1.747827
```

2.20 Lösung zur Aufgabe 1.3.5

```
df <- data.frame(Jahr=c(1:5),
                 Wirksamkeit=c(96, 84, 70, 58, 52)
)
```

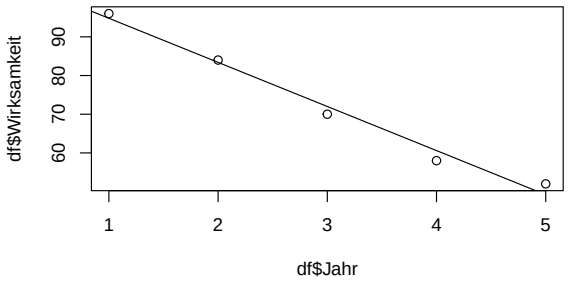
a) Führen Sie eine lineare Regression `Wirksamkeit erklärt durch Jahr` durch und plotten Sie Ihr Ergebnis.

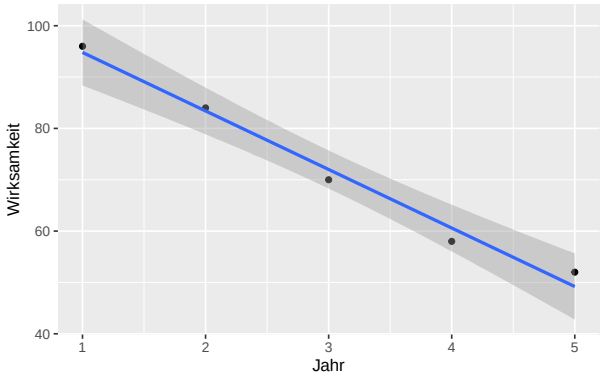
```
# Regression
fit <- lm(Wirksamkeit~Jahr, data=df)
summary(fit)
```

```
Call:
lm(formula = Wirksamkeit ~ Jahr, data = df)

Residuals:
   1    2    3    4    5
 1.2  0.6 -2.0 -2.6  2.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 106.2000     2.7350   38.83 3.76e-05 ***
Jahr        -11.4000     0.8246  -13.82 0.000819 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.608 on 3 degrees of freedom
Multiple R-squared:  0.9845,    Adjusted R-squared:  0.9794
F-statistic: 191.1 on 1 and 3 DF,  p-value: 0.0008192
```

```
# plot()
plot(df$Jahr, df$Wirksamkeit)
abline(fit)
# ggplot()
ggplot(df, aes(x=Jahr, y=Wirksamkeit)) +
  geom_point() +
  geom_smooth(method="lm")
```

```
`geom_smooth()` using formula = 'y ~ x'
```





b) Wie große ist der jährliche Wirksamkeitsverlust in %?

```
# Regression
fit$coefficients
```

```
(Intercept)        Jahr
      106.2       -11.4
```

Der Wirksamkeitsverlust beträgt 11.4% pro Jahr.

c) Nach wie vielen Jahren ist die Wirksamkeit bei 80%, und nach wie vielen bei 0%? Sind beide Werte gleich zuverlässig?

```
# anderes Modell
fit2 <- lm(Jahr ~ Wirksamkeit, data=df)
# 80% und 0%
predict(fit2, list(Wirksamkeit=c(80,0)))
```

```
       1        2
2.309091 9.218182
```

Nach 2.31 Jahren ist die Wirksamkeit bei 80%, nach 9.22 Jahren bei 0%.

2.21 Lösung zur Aufgabe 1.3.6

```
df <- data.frame(Dosis=c(2,2, 2,2,2,2,
                         3,3, 3,3,3,3,
                         3, 4,4,4,4,4, 4,4),
                 Tage =c(5,5, 6,6,6,6,
                         3,3, 5,5,5,5,
                         6, 3,3,3,3,3, 5,5))
```

a) Berechnen Sie die Regressionsgerade der Heilungstage in Abhängigkeit von der Dosis.

```
# Regression
fit <- lm(Tage~Dosis, data=df)
summary(fit)
```

```
Call:
lm(formula = Tage ~ Dosis, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6023 -0.5560  0.3513  0.3977  1.4440

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.7413     0.7941   9.749 1.32e-08 ***
Dosis        -1.0463     0.2517  -4.156 0.000593 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9059 on 18 degrees of freedom
Multiple R-squared:  0.4897,    Adjusted R-squared:  0.4614
F-statistic: 17.28 on 1 and 18 DF,  p-value: 0.000593
```

b) Berechnen Sie den Regressionskoeffizienten der Heilungstage in Abhängigkeit von der Dosis und interpretieren Sie ihn.

```
# Regression
fit$coefficients
```

```
(Intercept)       Dosis
   7.741313   -1.046332
```

Mit jeder Dosiserhöhung um 1 verkürzt sich die Heilungsdauer um ca. 1 Tag.

c) Berechnen Sie den Korrelationskoeffizienten und interpretieren Sie ihn.

```
# Regression
cor(df$Dosis, df$Tage)
```

```
[1] -0.69981
```

Der Korrelationskoeffizient ist größer als 0,5. Es liegt ein mittelstarker Zusammenhang vor.

d) Bestimmen Sie die erwartete Zeit, die für die Heilung mit einer Dosis von 5 mg benötigt wird. Ist diese Vorhersage zuverlässig? Begründen Sie die Antwort.

```
# Vorhersage
predict(fit, list(Dosis=5))
```

```
       1
2.509653
```

e) Welche Dosis muss angewendet werden, um in 4 Tagen zu heilen? Ist diese Vorhersage zuverlässig? Begründen Sie die Antwort.

```
# neues Modell
fit2 <- lm(Dosis~Tage, data=df)
# Vorhersage
predict(fit2, list(Tage=4))
```

```
       1
3.307427
```

2.22 Lösung zur Aufgabe 1.3.7

a) Laden Sie den Datensatz `heigths.weights.students` in Ihre R-Session.

```
# lade Datensatz
load(url("https://www.produnis.de/R/data/heights.weights.students.RData"))
```

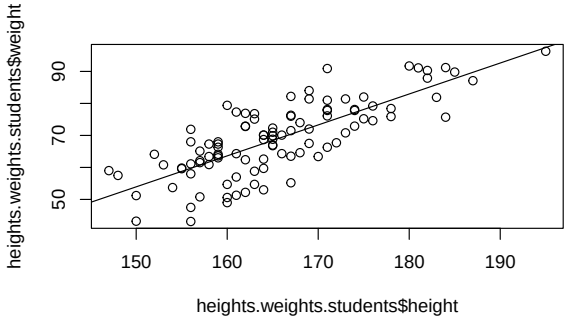
b) Führen Sie eine lineare Regression `Gewicht erklärt durch Größe` durch und plotten Sie Ihr Modell.

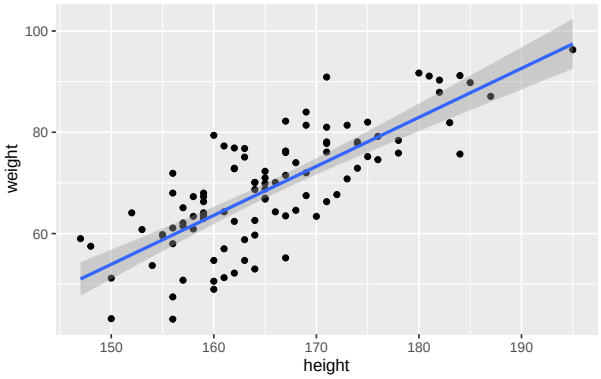
```
# Regression
fit <- lm(weight ~ height, data=heights.weights.students)
summary(fit)
```

```
Call:
lm(formula = weight ~ height, data = heights.weights.students)

Residuals:
     Min       1Q   Median       3Q      Max
-16.6372  -4.8272   0.9568   4.8008  16.6542

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -91.15252   13.28198  -6.863 6.16e-10 ***
height        0.96724    0.08009  12.077  < 2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.356 on 98 degrees of freedom
Multiple R-squared:  0.5981,    Adjusted R-squared:  0.594
F-statistic: 145.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

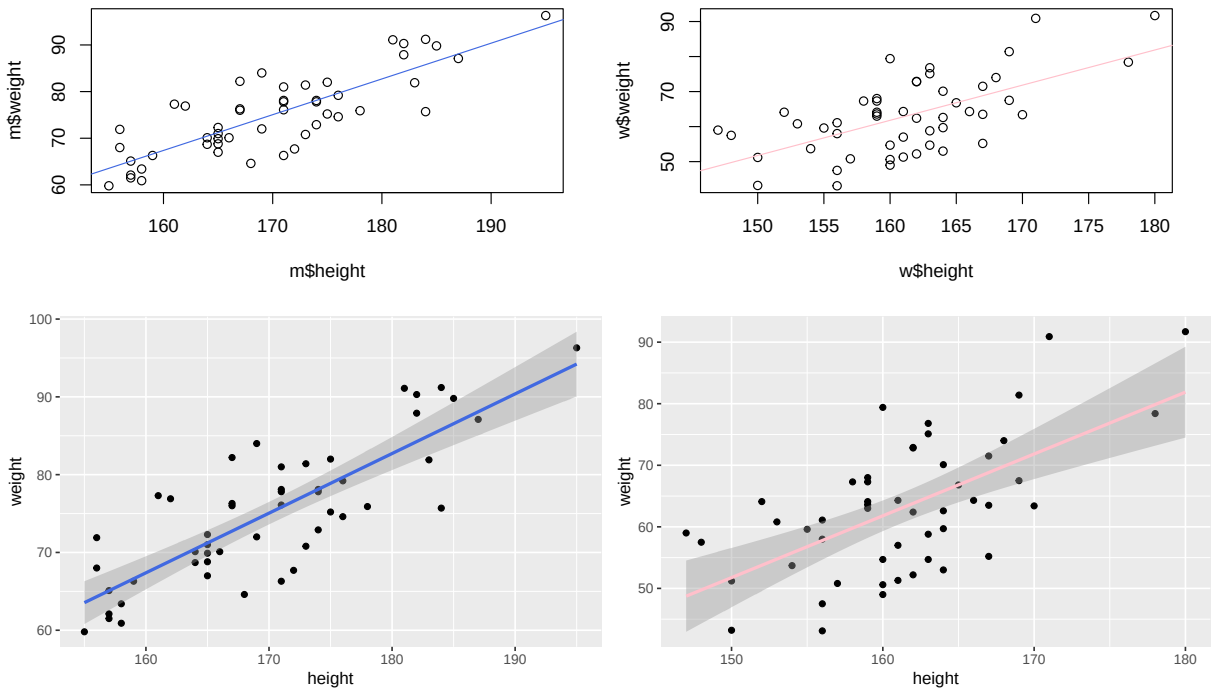
```r
# plot()
plot(heights.weights.students$height, heights.weights.students$weight)
abline(fit)
# ggplot()
ggplot(heights.weights.students, aes(x=height, y=weight)) +
  geom_point() +
  geom_smooth(method="lm")
```





c) Erstellen Sie eine Punktwolke inklusive Regressionsgeraden jeweils für Männer und Frauen getrennt.

```
m <- subset(heights.weights.students, sex=="male")
w <- subset(heights.weights.students, sex=="female")

fit1 <- lm(weight ~ height, data=m)
fit2 <- lm(weight ~ height, data=w)

## plot()
# männlich
plot(m$height, m$weight)
abline(fit1, col="royalblue")
# weiblich
plot(w$height, w$weight)
abline(fit2, col="pink")
## ggplot()
# männlich
ggplot(m, aes(x=height, y=weight))  +
  geom_point() +
  geom_smooth(method="lm", color="royalblue")
# weiblich
ggplot(w, aes(x=height, y=weight))  +
  geom_point() +
  geom_smooth(method="lm", color="pink")
```



d) Berechnen Sie die Bestimmtheitskoeffizienten (R^2) für beide Modelle. Welches Modell erklärt besser die Beziehung zwischen Gewicht und Größe, das der Männer oder das der Frauen? Begründen Sie die Antwort.

```
# Männer
summary(fit1)$r.squared
```

```
[1] 0.6699418
```

```
# Frauen
summary(fit2)$r.squared
```

```
[1] 0.3828876
```

Das Modell der Männer erklärt 0.67% der Streuung, und das der Frauen "nur" 0.38%. Somit ist das Modell für Männer *besser* als das der Frauen.

e) Was ist das zu erwartende Gewicht für einen Mann mit 170cm Körpergröße? Und für eine Frau der selben Größe?

```
# Männer
predict(fit1, list(height=170))
```

```
     1
75.048
```

```
# Frauen
predict(fit2, list(height=170))
```

```
      1
71.8338
```

2.23 Lösung zur Aufgabe 1.3.8

```
# lade Datensatz
load(url("https://www.produnis.de/R/data/neonates.RData"))
```

a) Erstellen Sie eine Kreuztabelle vom APGAR-Wert nach 1 Minute und dem Rauchverhalten der Mütter während der Schwangerschaft. Welche Schlüsse lassen sich ziehen?

```
# entweder
table(neonates$smoke, neonates$apgar1)
```

```
       2  3  4  5  6  7  8  9
  No   1  6 18 50 77 40 23  5
```

```
  Yes  3 15 20 31 20  6  5  0
```

```
# oder
xtabs(~ smoke + apgar1, data=neonates)
```

```
     apgar1
smoke  2  3  4  5  6  7  8  9
  No   1  6 18 50 77 40 23  5
  Yes  3 15 20 31 20  6  5  0
```

Kinder von Frauen, die nicht während der Schangerschaft rauchen, haben höhere APGAR1-Werte als Kinder von Raucherinnen.

b) Erstellen Sie eine Kreuztabelle vom APGAR-Wert nach 1 Minute und der Alterskategorie der Mütter. Welche Schlüsse lassen sich ziehen?

```
# entweder
table(neonates$age, neonates$apgar1)
```

```
                  2  3  4  5  6  7  8  9
  greater than 20  2 10 22 53 69 34 24  4
  less than 20     2 11 16 28 28 12  4  1
```

```
# oder
xtabs(~ age + apgar1, data=neonates)
```

```
                apgar1
age                2  3  4  5  6  7  8  9
  greater than 20  2 10 22 53 69 34 24  4
  less than 20     2 11 16 28 28 12  4  1
```

Kinder von Frauen, die älter als 20 Jahre sind, haben höhere APGAR1-Werte als Kinder von jüngeren Müttern.

c) Führen Sie eine lineare Regression für `Geburtsgewicht erklärt durch Anzahl täglich gerauchter Zigaretten` durch. Gibt es einen starken linearen Zusammenhang?

```
# Regression
fit <- lm(weight ~ cigarettes, data=neonates)
summary(fit)
```

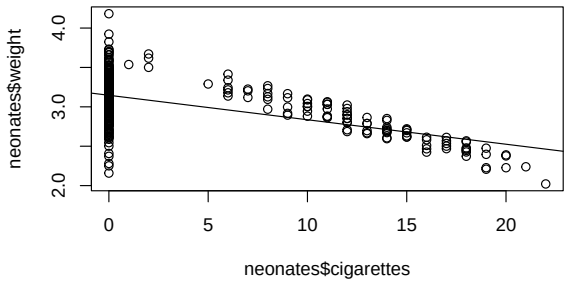
```
Call:
lm(formula = weight ~ cigarettes, data = neonates)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98756 -0.16656 -0.00649  0.18769  1.03544
```

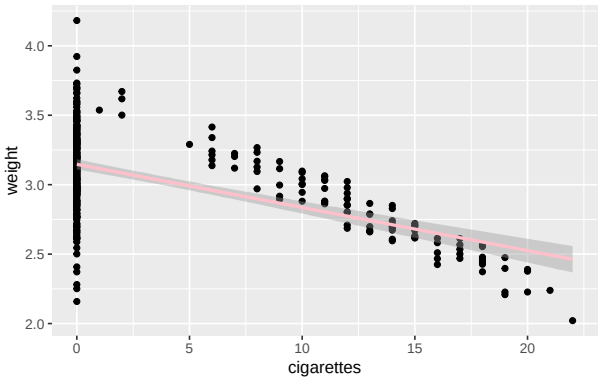
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.146557   0.018604  169.13   <2e-16 ***
cigarettes  -0.031067   0.002512  -12.37   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2832 on 318 degrees of freedom
Multiple R-squared:  0.3248,    Adjusted R-squared:  0.3227
F-statistic:   153 on 1 and 318 DF,  p-value: < 2.2e-16
```

Der Zusammenhang ist eher gering.

d) Plotten Sie Ihre Regression. Passt die Regressionsgerade gut zur Punktwolke?

```
# plot()
plot(neonates$cigarettes, neonates$weight)
abline(fit)
# ggplot()
ggplot(neonates, aes(x=cigarettes, y=weight))  +
  geom_point() +
  geom_smooth(method="lm", color="pink")
```





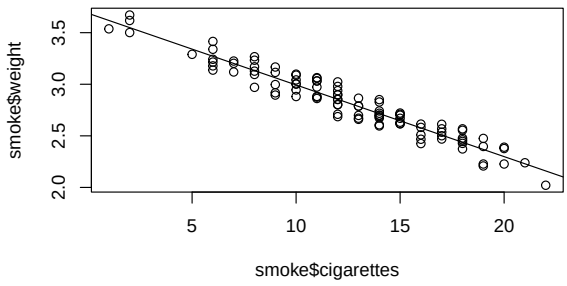
Der Zusammenhang wird durch die Nichtraucherinnen (0 Zigaretten) verzerrt.

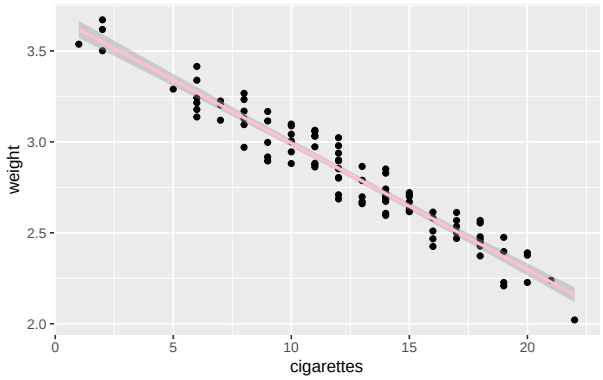
e) Wiederholen Sie die Regression, aber nutzen Sie dieses Mal nur Daten von Raucherinnen. Ist dieses Modell besser oder schlechter als das vorherige? Wieviel Gewicht verliert ein Neugeborenes nach diesem Modell pro täglich gerauchter Zigarette?

```
# Subset erzeugen
smoke <- subset(neonates, smoke=="Yes")
# Regression
fit2 <- lm(weight ~ cigarettes, data=smoke)
summary(fit2)
```

```
Call:
lm(formula = weight ~ cigarettes, data = smoke)

Residuals:
      Min        1Q    Median        3Q       Max
-0.168338 -0.057531  0.002855  0.068180  0.168662

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.687879   0.025542  144.38   <2e-16 ***
cigarettes  -0.069462   0.001928  -36.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08735 on 98 degrees of freedom
Multiple R-squared:  0.9298,    Adjusted R-squared:  0.9291
F-statistic:  1298 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
# plot()
plot(smoke$cigarettes, smoke$weight)
abline(fit2)
# ggplot()
ggplot(smoke, aes(x=cigarettes, y=weight))  +
  geom_point() +
  geom_smooth(method="lm", color="pink")
```





Der Zusammenhang ist nun sehr stark.

f) Welches Geburtsgewicht sagt dieses Modell für ein Neugeborenes vorher, dessen Mutter 5 Zigaretten täglich während der Schwangerschaft geraucht hat? Wieviel für eine Mutter, die 30 Zigaretten täglich raucht. Wie zuverlässich sind diese Ergebnisse?

```
# Vorhersage
predict(fit2, list(cigarettes=c(5, 30)))
```

```
       1        2
3.340570 1.604026
```

g) Ändert sich der lineare Zusammenhang, wenn die Daten nach Altersgruppen getrennt untersucht werden?

```
# Subset
s1 <- subset(smoke, age=="greater than 20")
s2 <- subset(smoke, age=="less than 20")

# neue Modelle
fit1 <- lm(weight ~ cigarettes, data=s1)
fit1 <- lm(weight ~ cigarettes, data=s2)

# vergleichen
summary(fit1)
```

```
Call:
lm(formula = weight ~ cigarettes, data = s2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.151567 -0.049334 -0.001749  0.062936  0.127103

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.65752    0.05636   64.90  < 2e-16 ***
cigarettes  -0.06833    0.00375  -18.22 6.33e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08286 on 20 degrees of freedom
Multiple R-squared:  0.9432,    Adjusted R-squared:  0.9404
F-statistic: 332.1 on 1 and 20 DF,  p-value: 6.333e-14
```

```
summary(fit2)
```

```
Call:
lm(formula = weight ~ cigarettes, data = smoke)

Residuals:
```

```
      Min        1Q    Median        3Q       Max
-0.168338 -0.057531  0.002855  0.068180  0.168662

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.687879   0.025542  144.38   <2e-16 ***
cigarettes  -0.069462   0.001928  -36.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08735 on 98 degrees of freedom
Multiple R-squared:  0.9298,    Adjusted R-squared:  0.9291
F-statistic:  1298 on 1 and 98 DF,  p-value: < 2.2e-16
```

Der Zusammenhang bleibt unabhängig von der Altersgruppe bestehen.

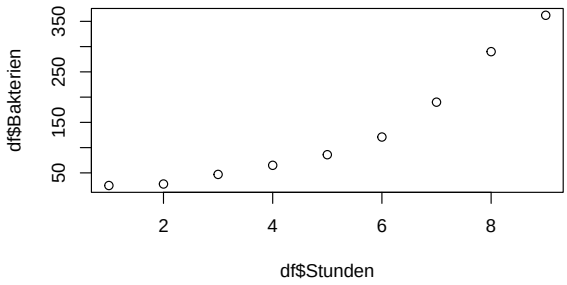
2.24 Lösung zur Aufgabe 1.4.1

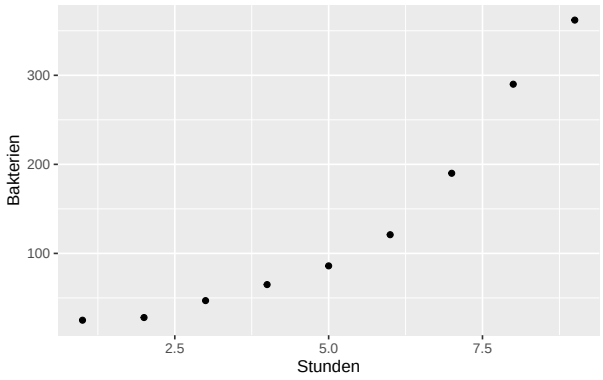
a) Erstellen Sie ein Datenframe mit den Variablen `Stunden` und `Bakterien`.

```
# Daten erzeugen
df <- data.frame(Stunden=c(1:9),
                 Bakterien=c(25, 28, 47, 65, 86, 121, 190, 290, 362))
```

b) Erzeugen Sie ein Scatterplot. Welche Regression würden Sie auf Grundlage des Plots vorschlagen?

```
# plot()
plot(df$Stunden, df$Bakterien)
# ggplot()
ggplot(df, aes(x=Stunden, y=Bakterien)) +
  geom_point()
```





Die Punktwolken sprechen für einen exponentiellen Anstieg.

c) Berechnen Sie die quadratischen und exponentiellen Modelle für die Bakterienvermehrung über die Zeit.

```
# quadratisch
q <- lm(Bakterien ~ Stunden + I(Stunden^2), data=df)
summary(q)
```

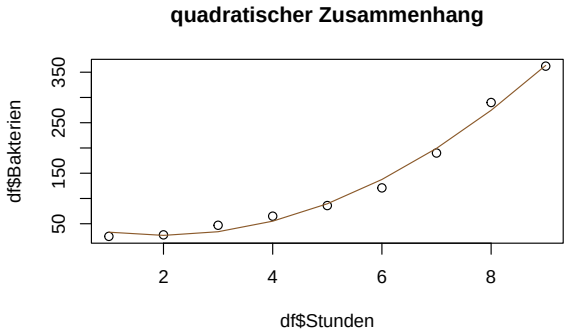
```
Call:
lm(formula = Bakterien ~ Stunden + I(Stunden^2), data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-16.617  -8.297  -1.430   9.916  15.442

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   53.2381    15.9862   3.330   0.0158 *
Stunden      -26.7420     7.3403  -3.643   0.0108 *
I(Stunden^2)   6.8009     0.7159   9.500 7.75e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.56 on 6 degrees of freedom
Multiple R-squared:  0.9919,    Adjusted R-squared:  0.9892
F-statistic: 368.8 on 2 and 6 DF,  p-value: 5.254e-07
```

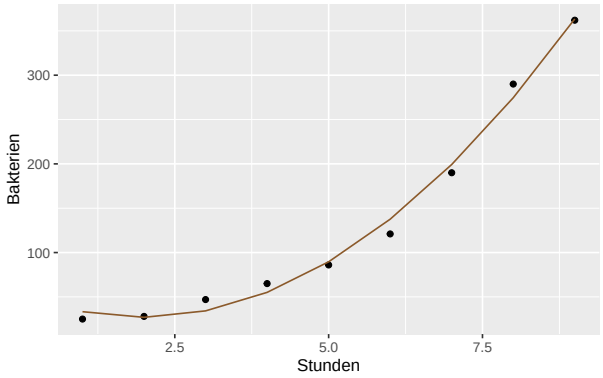
```
# exponentiell
e <- lm(log(Bakterien) ~ Stunden, data=df)
summary(e)
```

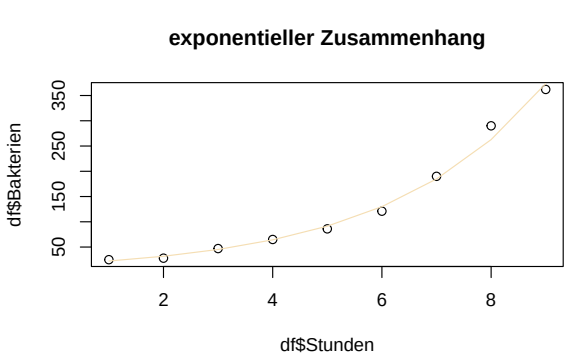
```
Call:
lm(formula = log(Bakterien) ~ Stunden, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.12676 -0.06057  0.01145  0.03920  0.11190

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.75498    0.06174   44.62 7.41e-10 ***
Stunden      0.35199    0.01097   32.08 7.39e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08498 on 7 degrees of freedom
Multiple R-squared:  0.9932,    Adjusted R-squared:  0.9923
F-statistic:  1029 on 1 and 7 DF,  p-value: 7.389e-09
```

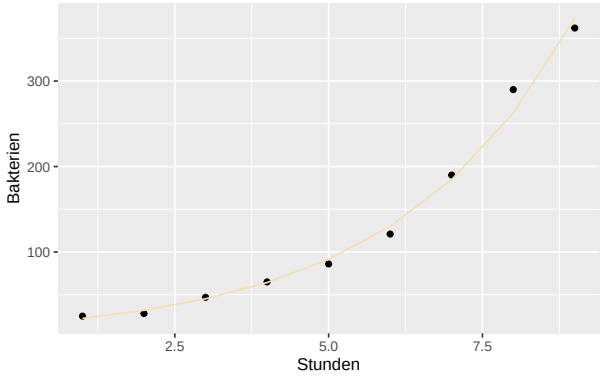
d) Plotten Sie das bessere Modell in die Punktwolke.

```
# Vorbereitung quadratisch
vorhersageQ <- predict(q, list(Stunden=df$Stunden))
# plot() quadratisch
plot(df$Stunden, df$Bakterien,
     main="quadratischer Zusammenhang")
lines(df$Stunden, vorhersageQ, col="tan4")
# ggplot() quadratisch
ggplot(df, aes(x=Stunden, y=Bakterien)) +
  geom_point() +
  geom_line(aes(y=vorhersageQ), col="tan4")
# Vorbereitung exponentiell
vorhersageE <- exp(predict(e, list(Stunden=df$Stunden)))

# plot() exponentiell
plot(df$Stunden, df$Bakterien,
     main="exponentieller Zusammenhang")
lines(df$Stunden, vorhersageE, col="wheat")
# ggplot() exponentiell
ggplot(df, aes(x=Stunden, y=Bakterien)) +
  geom_point() +
  geom_line(aes(y=vorhersageE), col="wheat")
```









e) Wie viele Bakterien werden nach dem besten Modell 3 Stunden nach Anlegen der Kultur vorhanden sein? Und nach 10 Stunden? Sind diese Vorhersagen zuverlässig?

```
# Vorhersage
exp(predict(e, list(Stunden=c(3, 10))))
```

```
        1         2
 45.19322 531.05241
```

Nach 3 Stunden können wir 46 Bakterien erwarten, nach 10 Stunden 532.

f) Machen Sie eine möglichst zuverlässige Vorhersage über die Zeit, die benötigt wird, um 100 Bakterien in der Kultur zu haben.

```
# neues Modell
df$BakterienLog <- log(df$Bakterien)
fit <- lm(BakterienLog ~ Stunden, data=df)
a <- exp(fit$coefficients[1])
b <- fit$coefficients[2]
# Vorhersage für 100 Bakterien
(log(100) - log(a)) / b
```

```
(Intercept)
   5.256395
```

Nach ca 5.3 Stunden sind 100 Bakterien zu erwarten.

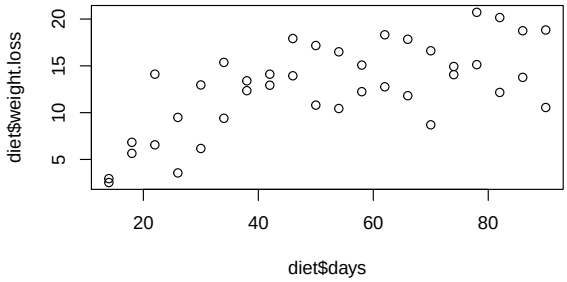
2.25 Lösung zur Aufgabe 1.4.2

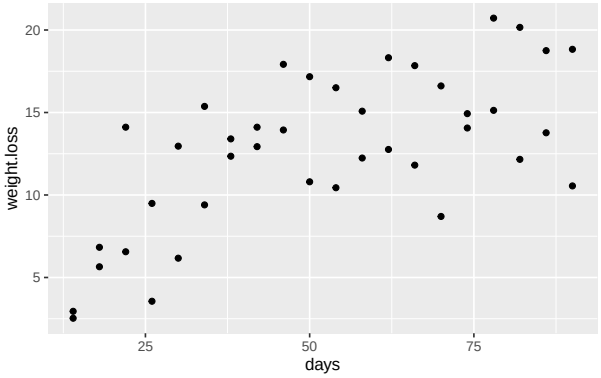
a) Laden Sie den Datensatz `diet` in Ihre R-Session.

```
# lade Datensatz
load(url("https://www.produnis.de/R/data/diet.RData"))
```

b) Erstellen Sie eine Punktwolke. Welche Art von Modell erklärt auf Grundlage der Punktwolke den Gewichtsverlust pro Diättag besser?

```
# plot()
plot(diet$days, diet$weight.loss)
# ggplot()
ggplot(diet, aes(x=days, y=weight.loss)) +
  geom_point()
```





c) Berechnen Sie das Regressionsmodell, welches den Gewichtsverlust mit der Anzahl an Diättagen am besten (im Vergleich zu anderen) erklären kann. Wird das Modell zuverlässige Vorhersagen machen?

```
# Vergleiche Bestimmheitsmaße verschiedener Modelle

# quadratisches Modell
q <- lm(weight.loss ~ days + I(days^2), data=diet)

# exponentielles Modell
e <- lm (log(weight.loss) ~ days, data=diet)

# logarithmisches Modell
l <- lm(weight.loss ~ log(days), data=diet)

# sigmoidales Modell
s <- lm(log(weight.loss) ~ I(1/days), data=diet)

result <- data.frame(Modell = c("quadratisch", "exponentiell",
                                "logarithmisch", "sigmoidal"),
                     R.square = c(summary(q)$r.square,
                                  summary(e)$r.square,
                                  summary(l)$r.square,
                                  summary(s)$r.square))
# Anzeigen
result[order(result$R.square, decreasing = TRUE),]
```

	Modell	R.square
4	sigmoidal	0.6662170
1	quadratisch	0.5397848
3	logarithmisch	0.5254856
2	exponentiell	0.4308936

Wir können für den Vergleich auch die Funktion `compare.lm()` aus dem `jgsbook`-Paket verwenden.

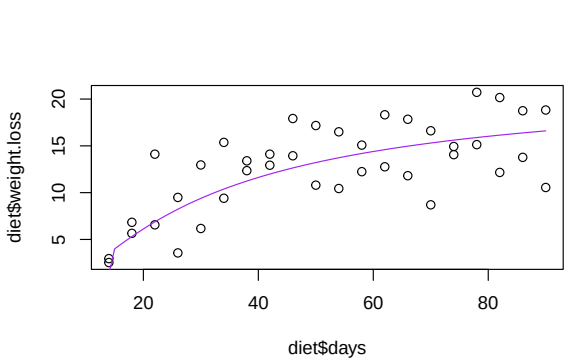
```
jgsbook::compare.lm(diet$weight.loss, diet$days)
```

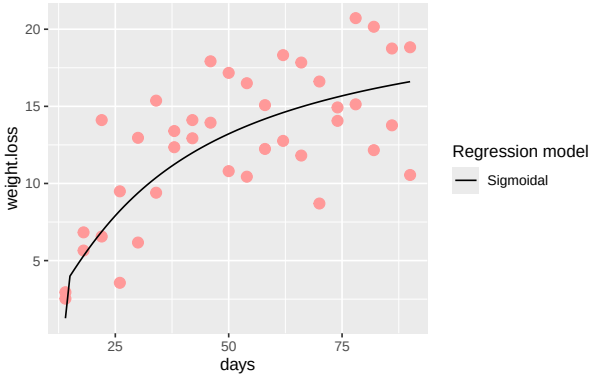
	Modell	R.square
6	sigmoidal	0.6662170
7	potenz	0.5684490
3	kubisch	0.5584355
2	quadratisch	0.5397848
5	logarithmisch	0.5254856
1	linear	0.4356390
4	exponentiell	0.4308936

Das sigmoidale Modell kann die Daten am besten erklären, da in diesem Modell das Bestimmtheitsmaß am größten ist.

d) Plotten Sie Ihr Modell.

```
# sigmoidales Modell
s <- lm(log(weight.loss) ~ I(1/days), data=diet)

# vorhersage vorbereiten
tage <- seq(min(diet$days), max(diet$days))
# alle "tage" vorhersagen
vorhersage <- predict(s, list(days=tage))
vorhersage[-1]=exp(vorhersage[-1])

# plot()
plot(diet$days, diet$weight.loss)
lines(tage, vorhersage, col="purple")
# ggplot()
# in Datenframe für ggplot speichern
helper <- data.frame(tage, vorhersage)

ggplot(diet, aes(x=days, y=weight.loss)) +
  geom_point(color="#FF9999", size=3) +
    # Legend
    scale_linetype("Regression model") +
    # Sigmoidal model
    geom_line(data=helper, aes(x=tage, y=vorhersage, linetype="Sigmoidal"))
```



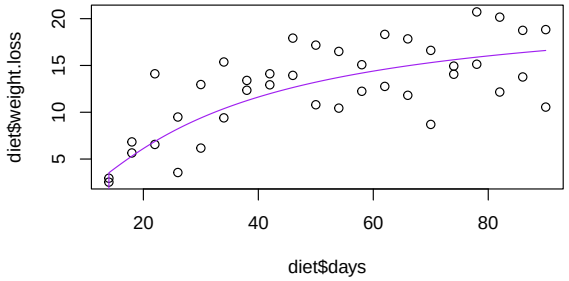


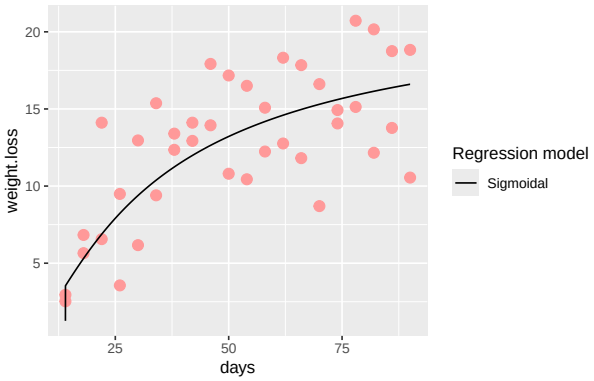
Auch die Vorhersagewerte für die Idealkurve können mittels `compare.lm()` und dem Parameter `predict=TRUE` erzeugt werden.

```
help <- jgsbook::compare.lm(diet$weight.loss, diet$days,
                            predict=TRUE)

# anschauen
head(help)
```

pred.x	line	quad	cube	expo	loga	sigm	power	logistic
14.00	7.645786	4.766305	3.526750	6.714088	5.166988	1.262199	4.966500	4.261378
14.01	7.647113	4.770032	3.532726	6.715040	5.171506	3.537806	4.969006	4.264669
14.02	7.648440	4.773757	3.538700	6.715992	5.176020	3.542430	4.971512	4.267962
14.03	7.649767	4.777482	3.544671	6.716944	5.180531	3.547052	4.974018	4.271256
14.04	7.651094	4.781207	3.550641	6.717896	5.185040	3.551674	4.976523	4.274552
14.05	7.652421	4.784931	3.556608	6.718848	5.189544	3.556296	4.979027	4.277850

```r
# plot()
plot(diet$days, diet$weight.loss)
lines(help$pred.x, help$sigm, col="purple")
# ggplot()
ggplot(diet, aes(x=days, y=weight.loss)) +
  geom_point(color="#FF9999", size=3) +
    # Legend
    scale_linetype("Regression model") +
    # Sigmoidal model
    geom_line(data=help, aes(x=pred.x, y=sigm, linetype="Sigmoidal"))
```





e) Berechnen Sie das Regressionsmodell, das den Gewichtsverlust anhand der Tage der Diät für die Gruppe der Personen, die sich nicht regelmäßig körperlich betätigen, am besten erklärt.

```
# Subset bilden
df1 <- subset(diet, exercise=="no")

# Vergleiche Bestimmheitsmaße verschiedener Modelle

# quadratisches Modell
q1 <- lm(weight.loss ~ days + I(days^2), data=df1)

# exponentielles Modell
e1 <- lm (log(weight.loss) ~ days, data=df1)

# logarithmisches Modell
l1 <- lm(weight.loss ~ log(days), data=df1)

# sigmoidales Modell
s1 <- lm(log(weight.loss) ~ I(1/days), data=df1)

result1 <- data.frame(Modell = c("quadratisch", "exponentiell",
                                 "logarithmisch", "sigmoidal"),
                      R.square = c(summary(q1)$r.square,
                                   summary(e1)$r.square,
                                   summary(l1)$r.square,
                                   summary(s1)$r.square))
result1[order(result1$R.square, decreasing = TRUE),]
```

	Modell	R.square
4	sigmoidal	0.7401212
1	quadratisch	0.7100610
3	logarithmisch	0.6494521
2	exponentiell	0.5222832

```
# oder mittels compare.lm()
jgsbook::compare.lm(df1$weight.loss, df1$days)
```

	Modell	R.square
6	sigmoidal	0.7401212
3	kubisch	0.7151929
2	quadratisch	0.7100610
7	potenz	0.6700051
5	logarithmisch	0.6494521
1	linear	0.5286338
4	exponentiell	0.5222832

Das sigmoidale Modell liefert wieder die beste Erklärung der Daten.

f) Wiederholen Sie die Analyse für die Gruppe, die sich regelmäßig körperlich betätigt.

```
# Subset bilden
df2 <- subset(diet, exercise=="yes")

# Vergleiche Bestimmheitsmaße verschiedener Modelle

# quadratisches Modell
q2 <- lm(weight.loss ~ days + I(days^2), data=df2)

# exponentielles Modell
e2 <- lm (log(weight.loss) ~ days, data=df2)

# logarithmisches Modell
l2 <- lm(weight.loss ~ log(days), data=df2)

# sigmoidales Modell
s2 <- lm(log(weight.loss) ~ I(1/days), data=df2)

result2 <- data.frame(Modell = c("quadratisch", "exponentiell",
                                 "logarithmisch", "sigmoidal"),
                      R.square = c(summary(q2)$r.square,
                                   summary(e2)$r.square,
                                   summary(l2)$r.square,
                                   summary(s2)$r.square))
result2[order(result1$R.square, decreasing = TRUE),]
```

	Modell	R.square
4	sigmoidal	0.8305013
1	quadratisch	0.7791671
3	logarithmisch	0.7885173

2	exponentiell	0.4945564

```
# oder mittels compare.lm()
jgsbook::compare.lm(df2$weight.loss, df2$days)
```

	Modell	R.square
3	kubisch	0.8326179
6	sigmoidal	0.8305013
5	logarithmisch	0.7885173
2	quadratisch	0.7791671
7	potenz	0.6704843
1	linear	0.6623502
4	exponentiell	0.4945564

Das kubische Modell liefert hier die beste Erklärung der Daten.

g) Benutzen Sie die erstellen Modelle, um den Gewichtsverlust nach 30 und nach 100 Tagen Diät für Personen, die sich körperlich betätigen, und für solche, die dies nicht tun, vorherzusagen. Sind diese Vorhersagen zuverlässig?

```
# Vorhersage Kein Sport
exp(predict(s1, list(days=c(30,100))))
```

```
       1        2
 7.808926 13.806339
```

```
# Vorhersage Sport
exp(predict(s2, list(days=c(30,100))))
```

```
       1        2
11.28578 21.13143
```

2.26 Lösung zur Aufgabe 1.4.3

```
df <- data.frame(Stunden = c(2:8),
                 Konzentration = c(25, 36, 48, 64, 86, 114, 168))
```

a) Benutzen Sie ein exponentielles Modell, um die Konzentration nach 10 Stunden vorherzusagen. Ist die Vorhersage zuverlässig?

```
# exponentielles Modell
fit <- lm(log(Konzentration) ~ Stunden, data=df)
summary(fit)
```

```
Call:
lm(formula = log(Konzentration) ~ Stunden, data = df)

Residuals:
        1         2         3         4         5         6         7
-0.023147  0.034218  0.014623 -0.004972 -0.016786 -0.042212  0.038276

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.627468   0.033673   78.03 6.55e-09 ***
Stunden     0.307277   0.006253   49.14 6.59e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03309 on 5 degrees of freedom
Multiple R-squared:  0.9979,    Adjusted R-squared:  0.9975
F-statistic:  2415 on 1 and 5 DF,  p-value: 6.594e-08
```

```
# Vorhersage 10 Stunden
exp(predict(fit, list(Stunden=10)))
```

```
     1
298.94
```

Nach 10 Stunden beträgt die Konzentration 298,94 mg/dl.
Das Bestimmtheitsmaß R^2 des Modells ist mit 0,9979 sehr groß. Die Daten werden sehr gut durch das Modell erklärt.

b) Benutzen Sie ein logarithmisches Modell um zu bestimmen, nach wie vielen Stunden eine Konzentration von 100 mg/dl erreicht sein wird.

```
# exponentielles Modell
fit <- lm(log(Konzentration) ~ Stunden, data=df)

# coefficienten
a <- fit$coefficient[1]
b <- fit$coefficient[2]

# Vorhersage 100 mg/dl
( log(100) - a ) / b
```

```
(Intercept)
   6.436209
```

Nach 6,436209 Stunden sind 100 mg/dl erreicht.

2.27 Lösung zur Aufgabe 1.5.1

a) Lassen Sie in R eine beliebige Poker-Spielkarte ziehen.

```
# lade Kartenspiele
load(url("https://www.produnis.de/R/data/cards.RData"))
```

```
# ziehe Karte
sample(poker, 1)
```

```
[1] Pik As
52 Levels: Kreuz 2 < Karo 2 < Herz 2 < Pik 2 < Kreuz 3 < Karo 3 < ... < Pik As
```

```
# alternativ kann das 'probs'-Paket verwendet werden
karten <- probs::cards(makespace=TRUE)
probs::sim(karten, ntrials=1)
```

rank	suit
J	Club

b) Lassen Sie in R 2 Münzen werfen.

```
# lade Datensatz
coin <- c("Kopf", "Zahl")

# wirf 2 Münzen
sample(coin, 2, replace=TRUE)
```

```
[1] "Kopf" "Zahl"
```

```
# alternativ kann das 'probs'-Paket verwendet werden
coin <- probs::tosscoin(2, makespace=TRUE)
probs::sim(coin, ntrials=1)
```

toss1	toss2
H	T

c) Lassen Sie in R 2 Würfel werfen.

```
# lade Datensatz
würfel <- c(1:6)

# wirf 2 Münzen
sample(würfel, 2, replace=TRUE)
```

```
[1] 2 4
```

```
# alternativ kann das 'probs'-Paket verwendet werden
würfel <- probs::rolldie(2, makespace=TRUE)
probs::sim(würfel, ntrials=1)
```

X1	X2
5	4

2.28 Lösung zur Aufgabe 1.5.2

a) Wiederholen Sie die Zufallsexperimente und lassen Sie R 10 mal, 100 mal 1.000 mal und 1.000.000 mal zwei Münzen werfen. Erstellen Sie je eine relative Häufigkeitstabelle der Ergebnisse. Wie sind die Tabellen zu bewerten?

```
# erzeuge Wahrscheinlichkeitsraum
münzen <- probs::tosscoin(2, makespace=TRUE)

# werfe 10mal 2 Münzen
versuch <- probs::sim(münzen, ntrials=10)
# berechne Wahrscheinlichkeitsverteilung
probs::empirical(versuch)
```

toss1	toss2	probs
T	H	0.3
H	T	0.4
T	T	0.3

```
# werfe 100mal 2 Münzen
versuch <- probs::sim(münzen, ntrials=100)
# berechne Wahrscheinlichkeitsverteilung
probs::empirical(versuch)
```

toss1	toss2	probs
H	H	0.30
T	H	0.20
H	T	0.28
T	T	0.22

```
# werfe 1.000mal 2 Münzen
versuch <- probs::sim(münzen, ntrials=1000)
# berechne Wahrscheinlichkeitsverteilung
probs::empirical(versuch)
```

toss1	toss2	probs
H	H	0.252
T	H	0.248
H	T	0.249
T	T	0.251

```
# werfe 1.000.000mal 2 Münzen
versuch <- probs::sim(münzen, ntrials=1000000)
# berechne Wahrscheinlichkeitsverteilung
probs::empirical(versuch)
```

toss1	toss2	probs
H	H	0.249949
T	H	0.249696
H	T	0.250325
T	T	0.250030

Mit zunehmender Wiederholung nähern sich die Wahrscheinlichkeitsverteilungen den relativen Häufigkeiten an.

b) Welche theoretischen Wahrscheinlichkeiten haben die möglichen Wurfergebnisse? Stimmen diese mit den beobachteten Ergebnissen überein?

Je häufiger das Zufallsexperiment wiederholt wird, desto mehr nähern sich die beobachteten Wahrscheinlichkeiten den theoretischen Wahrscheinlichkeiten an.

2.29 Lösung zur Aufgabe 1.5.3

a) Ziehen Sie zufällig 3 Boxen, ohne zurücklegen.

```
boxen <- c("A", "A", "A", "B", "B", "C")

# ziehe 3 Boxen ohne Zurücklegen
sample(boxen, 3, replace=FALSE)
```

```
[1] "B" "A" "B"
```

Ziehen Sie zufällig 3 Boxen, diesmal mit zurücklegen.

```
boxen <- c("A", "A", "A", "B", "B", "C")

# ziehe 3 Boxen ohne Zurücklegen
sample(boxen, 3, replace=TRUE)
```

```
[1] "B" "A" "B"
```

2.30 Lösung zur Aufgabe 1.5.4

a) Erstellen Sie ein Datenframe mit den Variablen `Windpocken`, `Masern`, `Röteln` und `Häufigkeit` und übertragen Sie die Daten.

```
df <- tribble(
  ~Windpocken, ~Masern, ~Röteln, ~Häufigkeit,
  "No",        "No",    "No",    2654,
  "No",        "No",    "Yes",   1436,
  "No",        "Yes",   "No",    1682,
  "No",        "Yes",   "Yes",   668,
  "Yes",       "No",    "No",    1747,
  "Yes",       "No",    "Yes",   476,
  "Yes",       "Yes",   "No",    876,
  "Yes",       "Yes",   "Yes",   265
)
```

b) Erstellen Sie den Wahrscheinlichkeitsraum der Lebenszeitprävalenz.

```
# Wahrscheinlichkeitsraum
wr <- probs::probspace(df[,-4], probs=df$Häufigkeit/sum(df$Häufigkeit))
wr
```

Windpocken	Masern	Röteln	probs
No	No	No	0.2707058
No	No	Yes	0.1464708
No	Yes	No	0.1715626
No	Yes	Yes	0.0681355
Yes	No	No	0.1781926
Yes	No	Yes	0.0485516
Yes	Yes	No	0.0893513
Yes	Yes	Yes	0.0270298

```
# Erstelle daraus die marginale Verteilung
probs::marginal(wr)
```

Windpocken	Masern	Röteln	probs
No	No	No	0.2707058
Yes	No	No	0.1781926
No	Yes	No	0.1715626
Yes	Yes	No	0.0893513
No	No	Yes	0.1464708
Yes	No	Yes	0.0485516
No	Yes	Yes	0.0681355
Yes	Yes	Yes	0.0270298

c) Wie groß ist die Wahrscheinlichkeit, dass eine zufällig gezogene Person Windpocken hatte?

```
# Wahrscheinlichkeitsraum
wr <- probs::probspace(df[,-4], probs=df$Häufigkeit/sum(df$Häufigkeit))
# berechne Wahrscheinlichkeit
probs::Prob(wr, event=Windpocken=="Yes")
```

```
[1] 0.3431253
```

Die Wahrscheinlichkeit beträgt 34.31%.

d) Wie groß ist die Wahrscheinlichkeit, dass eine zufällig gezogene Person Windpocken oder Masern hatte?

```
# berechne Wahrscheinlichkeit
probs::Prob(wr, event=Windpocken=="Yes" | Röteln=="Yes")
```

```
[1] 0.5577315
```

Die Wahrscheinlichkeit beträgt 55.77%.

e) Wie groß ist die Wahrscheinlichkeit, dass eine zufällig gezogene Person Masern und Röteln hatte?

```
# berechne Wahrscheinlichkeit
probs::Prob(wr, event=Masern=="Yes" & Röteln=="Yes")
```

```
[1] 0.09516524
```

Die Wahrscheinlichkeit beträgt 9.52%.

f) Wie groß ist die Wahrscheinlichkeit, dass eine zufällig gezogene Person, die bereits an Masern erkrankte, nun an Windpocken erkrankt?

```
# berechne Wahrscheinlichkeit
probs::Prob(wr, event=Windpocken=="Yes", given= Masern=="Yes")
```

```
[1] 0.3268404
```

Die Wahrscheinlichkeit beträgt 32.68%.

g) Wie groß ist die Wahrscheinlichkeit, dass eine zufällig gezogene Person, die keine Masern und keine Röteln hatte, an Windpocken erkrankt?

```
# berechne Wahrscheinlichkeit
probs::Prob(wr, event=Masern=="No" & Röteln=="No",
                given= Windpocken=="Yes")
```

```
[1] 0.5193222
```

Die Wahrscheinlichkeit beträgt 51.93%.

2.31 Lösung zur Aufgabe 1.5.5

a) Erstellen Sie ein Datenframe mit den Variablen `Schwanger`, `Testergebnis` und `Häufigkeit`.

```
df <- tribble(
  ~Schwanger, ~Test, ~Häufigkeit,
    "Nein", "-",   3876,
    "Nein", "+",  47,
    "Ja", "-",  12,
    "Ja", "+", 131
)
```

b) Erstellen Sie den Wahrscheinlichkeitsraum.

```
# Wahrscheinlichkeitsraum
wr <- probs::probspace(df[,-3], probs=df$Häufigkeit/sum(df$Häufigkeit))
wr
```

Schwanger	Test	probs
Nein	-	0.9532710
Nein	+	0.0115593
Ja	-	0.0029513
Ja	+	0.0322184

```
# Erstelle daraus die marginale Verteilung
probs::marginal(wr)
```

Schwanger	Test	probs
Ja	-	0.0029513
Nein	-	0.9532710
Ja	+	0.0322184
Nein	+	0.0115593

c) Berechnen Sie die Prävalenz der Schwangerschaften.

```
# Wahrscheinlichkeitsraum
probs::Prob(wr, event=Schwanger=="Ja")
```

```
[1] 0.0351697
```

Die Prävalenz liegt bei 3.52%.

d) Wie groß ist die Wahrscheinlichkeit, ein positives Testergebnis zu ziehen?

```
# Wahrscheinlichkeitsraum
probs::Prob(wr, event=Test=="+")
```

```
[1] 0.04377767
```

Die Wahrscheinlichkeit liegt bei 4.38%.

e) Bestimmen Sie die Sensitivität des Tests

```
# Wahrscheinlichkeitsraum
probs::Prob(wr, event=Test=="+", given= Schwanger=="Ja")
```

```
[1] 0.9160839
```

Die Sensitivität liegt bei 91.61%.

f) Bestimmen Sie die Spezifität des Tests

```
# Wahrscheinlichkeitsraum
probs::Prob(wr, event=Test=="-", given= Schwanger=="Nein")
```

```
[1] 0.9880194
```

Die Spezifität liegt bei 98.8%.

g) Bestimmen Sie den positiv prädiktiven Wert des Tests

```
# Wahrscheinlichkeitsraum
probs::Prob(wr, event=Schwanger=="Ja", given=Test=="+")
```

```
[1] 0.7359551
```

Der positiv prädiktive Wert liegt bei 73.6%.

h) Bestimmen Sie den negativ prädiktiven Wert des Tests

```
# Wahrscheinlichkeitsraum
probs::Prob(wr, event=Schwanger=="Nein", given=Test=="-")
```

```
[1] 0.9969136
```

Der negative prädiktive Wert liegt bei 99.69%.

Alternativ kann auch die Funktion `sens.spec()` aus dem Paket `jgsbook` verwendet werden:

```
jgsbook::sens.spec(rp=131, fp=12, rn=3876, fn=47)
```

sens	spec	ppw	npw
73.6	99.69	91.61	98.8

2.32 Lösung zur Aufgabe 1.5.6

Erstelle den Ereignisraum des Zufallsexperiments, das aus dem Werfen einer Münze, dem Werfen eines Würfels und dem Ziehen einer Karte aus einem französischen Kartenspiel besteht.

```
würfel <- 1:6
münze <- c("Kopf", "Zahl")
bild <- c(7:10 ,"B", "D", "K", "A")
farbe <- c("Kreuz", "Pik", "Karo", "Herz")

Ereignisraum <- expand.grid(Münze=münze, Bild=bild,
                            Farbe=farbe, Würfel=würfel)
head(Ereignisraum)
```

Münze	Bild	Farbe	Würfel
Kopf	7	Kreuz	1
Zahl	7	Kreuz	1
Kopf	8	Kreuz	1
Zahl	8	Kreuz	1
Kopf	9	Kreuz	1
Zahl	9	Kreuz	1

2.33 Lösung zur Aufgabe 1.5.7

```
df <- tribble(
  ~Impfung, ~Grippe, ~Häufigkeit,
   "Nein",   "Nein",   418,
   "Nein",    "Ja",    312,
   "Ja",     "Nein",   233,
   "Ja",     "Ja",    37)
```

a) Erzeugen Sie den Wahrscheinlichkeitsraum

```
wr <- probs::probspace(df[,-3],probs=df$Häufigkeit/sum(df$Häufigkeit))
wr
```

Impfung	Grippe	probs
Nein	Nein	0.418
Nein	Ja	0.312
Ja	Nein	0.233
Ja	Ja	0.037

```
# Erstelle daraus die marginale Verteilung
probs::marginal(wr)
```

Impfung	Grippe	probs
Ja	Ja	0.037
Nein	Ja	0.312
Ja	Nein	0.233
Nein	Nein	0.418

b) Wie groß ist die Wahrscheinlichkeit, dass eine zufällig ausgewählte Person geimpft ist?

```
probs::Prob(wr,event=Impfung=="Ja")
```

```
[1] 0.27
```

Die Wahrscheinlichkeit beträgt 27%.

c) Wie hoch ist die Prävalenz der Grippe?

```
probs::Prob(wr,event=Grippe=="Ja")
```

```
[1] 0.349
```

Die Prävalenz beträgt 34.9%.

d) Wie groß ist die Wahrscheinlichkeit, dass geimpfte Personen an Grippe erkranken? Ist die Impfung effektiv?

```
probs::Prob(wr,event=Grippe=="Ja", given=Impfung=="Ja")
```

```
[1] 0.137037
```

Die Wahrscheinlichkeit beträgt 13.7%.

2.34 Lösung zur Aufgabe 1.5.8

```
df <- tribble(
  ~Ebola, ~Test, ~Häufigkeit,
   "Nein",    "+",    28,
   "Nein",     "-",    97465,
   "Ja",      "+",    147,
   "Ja",      "-",    65)
```

a) Erzeugen Sie den Wahrscheinlichkeitsraum

```
wr <- probs::probspace(df[,-3],probs=df$Häufigkeit/sum(df$Häufigkeit))
wr
```

Ebola	Test	probs
Nein	+	0.0002866
Nein	-	0.9975436
Ja	+	0.0015045
Ja	-	0.0006653

```
# Erstelle daraus die marginale Verteilung
probs::marginal(wr)
```

Ebola	Test	probs
Ja	-	0.0006653
Nein	-	0.9975436
Ja	+	0.0015045
Nein	+	0.0002866

b) Berechnen Sie die Prävalenz von Ebola in der Bevölkerung.

```
probs::Prob(wr,event=Ebola=="Ja")
```

```
[1] 0.002169797
```

Die Prävalenz beträgt 0.22%.

c) Wie hoch ist die Wahrscheinlichkeit, ein negatives Testergebnis zu erhalten?

```
probs::Prob(wr,event=Test=="-")
```

```
[1] 0.9982089
```

Die Prävalenz beträgt 99.82%.

d) Berechnen Sie die Sensitivität und Spezifität des Tests.

```
# entweder
# Sensitivität
probs::Prob(wr,event=Test=="+", given=Ebola=="Ja")
```

```
[1] 0.6933962
```

```
# Spezifität
probs::Prob(wr,event=Test=="-", given=Ebola=="Nein")
```

```
[1] 0.9997128
```

```
# oder
jgsbook::sens.spec(fp=28, rn=97465, rp=147, fn=65)
```

sens	spec	ppw	npw
69.34	99.97	84	99.93

e) Kann der Test besser Erkrankte erkennen, oder Gesunde?

```
jgsbook::sens.spec(fp=28, rn=97465, rp=147, fn=65)
```

sens	spec	ppw	npw
69.34	99.97	84	99.93

Er kann besser Gesunde erkennen.

f) Wenn eine Person einen positiven Test erhält, wie hoch ist dann die Wahrscheinlichkeit, dass er tatsächlich krank ist?

```
# positiv prädiktiv
probs::Prob(wr,event=Ebola=="Ja", given=Test=="+")
```

```
[1] 0.84
```

Die Wahrscheinlichkeit liegt bei 84%.

g) Wenn eine Person einen negativen Test erhält, wie hoch ist dann die Wahrscheinlichkeit, dass er tatsächlich gesund ist?

```
# negativ prädiktiv
probs::Prob(wr,event=Ebola=="Nein", given=Test=="-")
```

```
[1] 0.9993335
```

Die Wahrscheinlichkeit liegt bei 99.93%.

2.35 Lösung zur Aufgabe 1.6.1

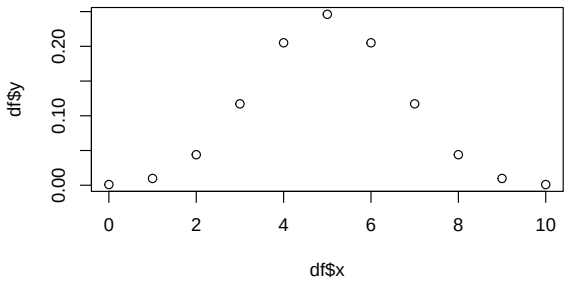
a) Berechnen Sie die Wahrscheinlichkeitsverteilung von `X`

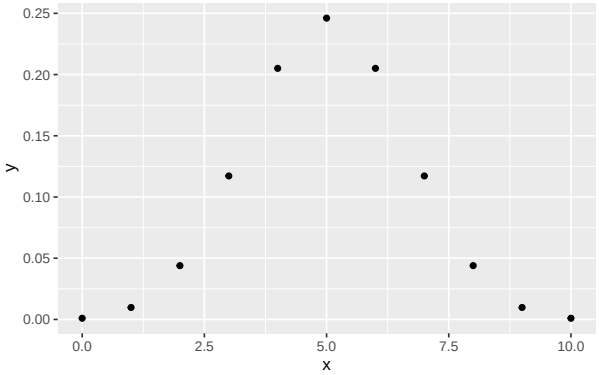
```
# mögliche Ausprägungen von x
x <- 0:10
# Wahrscheinlichkeiten berechnen
w <- dbinom(x, size = 10, prob = 0.50)

data.frame(Kopf=x, Wahrscheinlichkeit = w)
```

Kopf	Wahrscheinlichkeit
0	0.0009766
1	0.0097656
2	0.0439453
3	0.1171875
4	0.2050781
5	0.2460938
6	0.2050781
7	0.1171875
8	0.0439453
9	0.0097656
10	0.0009766

b) Plotten Sie die Wahrscheinlichkeitsfunktion von `X`

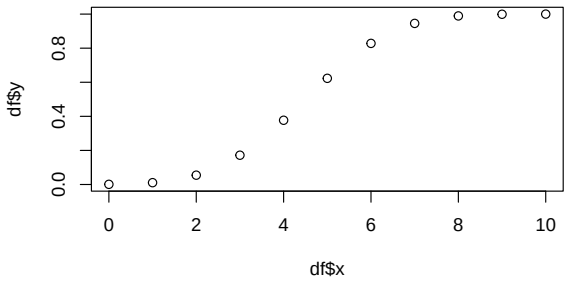
```
# mögliche Ausprägungen von x
x <- 0:10
df = data.frame(x, y=dbinom(x, size = 10, prob = 0.50))

# plot()
plot(df$x, df$y)
# ggplot()
ggplot(df, aes(x=x, y=y)) +
  geom_point()
```

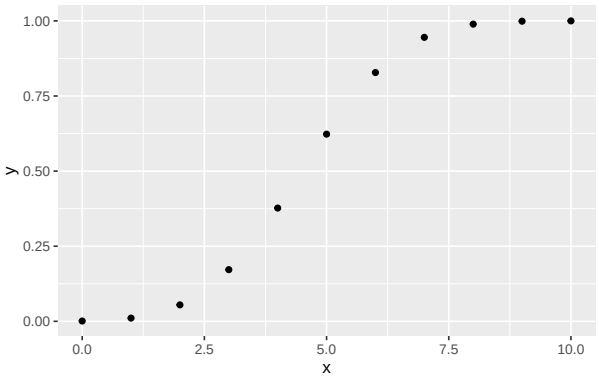




c) Plotten Sie die Dichteverteilung

```
# mögliche Ausprägungen von x
x <- 0:10
df = data.frame(x, y=pbinom(x, size = 10, prob = 0.50))

# plot()
plot(df$x, df$y)
# ggplot()
ggplot(df, aes(x=x, y=y)) +
  geom_point()
```





d) Berechnen Sie die Wahrscheinlichkeit, 7 mal `Kopf` zu werfen.

```
dbinom(7, size = 10, prob = 0.50)
```

```
[1] 0.1171875
```

e) Berechnen Sie die Wahrscheinlichkeit, weniger als als 4 mal `Kopf` zu werfen.

```
pbinom(4, size = 10, prob = 0.50, lower.tail=TRUE)
```

```
[1] 0.3769531
```

f) Berechnen Sie die Wahrscheinlichkeit, mehr als als 5 mal `Kopf` zu werfen.

```
pbinom(5, size = 10, prob = 0.50, lower.tail=FALSE)
```

```
[1] 0.3769531
```

g) Berechnen Sie die Wahrscheinlichkeit, 2 bis 8 mal `Kopf` zu werfen.

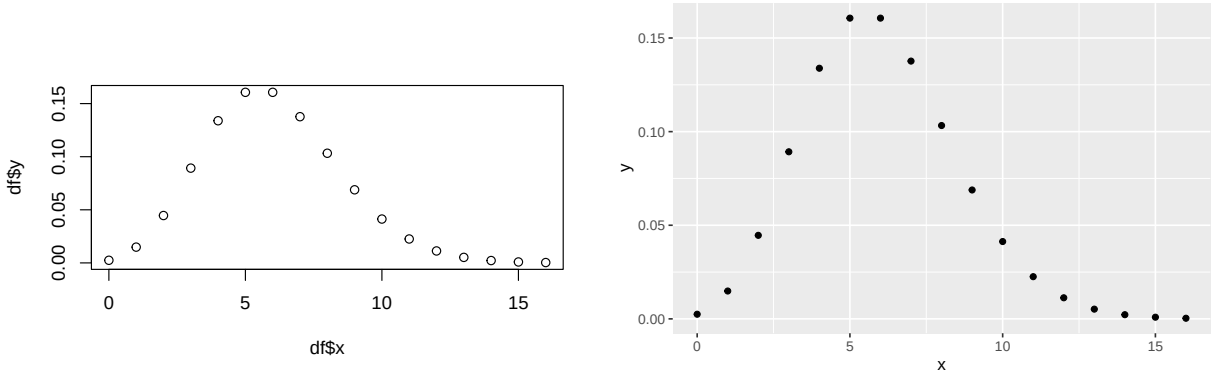
```
# weniger als 2mal
w2 <- pbinom(1, size = 10, prob = 0.5)
# weniger als 8mal
w8 <- pbinom(8, size = 10, prob = 0.5)

# Wahrscheinlichkeit 2 bis 8 Köpfe zu werfen
w8 - w2
```

```
[1] 0.9785156
```

2.36 Lösung zur Aufgabe 1.6.2

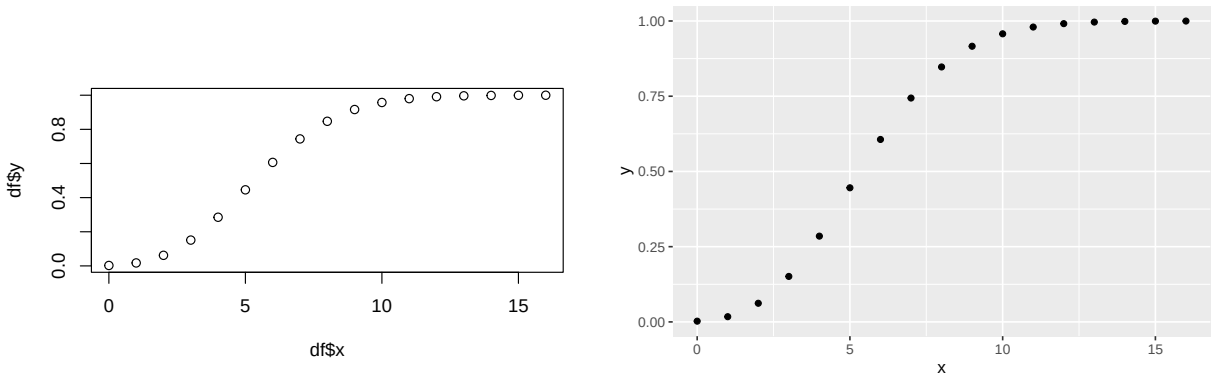
a) Plotten Sie die Wahrscheinlichkeitsfunktion von `X`

```
#  x-Achse
x = 0:16
# Poisson-Werte
df = data.frame(x, y=dpois(x, lambda = 6))

# plot()
plot(df$x, df$y)
# ggplot()
ggplot(df, aes(x=x, y=y)) +
  geom_point()
```



b) Plotten Sie die Verteilungsfunktion von `X`

```
#  x-Achse
x = 0:16
# Poisson-Werte
df = data.frame(x, y=ppois(x, lambda = 6))

# plot()
plot(df$x, df$y)
# ggplot()
ggplot(df, aes(x=x, y=y)) +
  geom_point()
```



c) Wie groß ist die Wahrscheinlicheit, dass an einem zufälligen Tag (nur) 1 Geburt stattfindet?

```
dpois(1, lambda = 6)
```

```
[1] 0.01487251
```

d) Wie groß ist die Wahrscheinlicheit, dass an einem zufälligen Tag weniger als 6 Geburten stattfinden?

```
ppois(5, lambda = 6, lower.tail=TRUE)
```

```
[1] 0.4456796
```

e) Wie groß ist die Wahrscheinlicheit, dass an einem zufälligen Tag 4 oder mehr Geburten stattfinden?

```
ppois(3, lambda = 6, lower.tail=FALSE)
```

```
[1] 0.8487961
```

f) Wie groß ist die Wahrscheinlicheit, dass an einem zufälligen Tag 4 bis 8 Geburten stattfinden?

```
# entweder
sum(dpois(4:8, lambda=6))
```

```
[1] 0.6960336
```

```
# oder
ppois(8, lambda=6) - ppois(3, lambda=6)
```

```
[1] 0.6960336
```

g) Wie groß ist die Wahrscheinlicheit, dass in einer Woche zwischen 30 und 40 Geburten stattfinden?

```
# lambda für eine Woche = 7* 6 = 42
ppois(40, lambda=42) - ppois(29, lambda=42)
```

```
[1] 0.3959028
```

2.37 Lösung zur Aufgabe 1.6.3

a) berechnen Sie die Wahrscheinlichkeitsverteilung des binomialen Modells $B(30, 0.1)$.

```
dbinom(c(0,1,2,3,4,5,6,7,8,9,10), size = 30, prob = 0.1)
```

```
 [1] 0.0423911583 0.1413038609 0.2276562204 0.2360879322 0.1770659492
 [6] 0.1023047706 0.0473633197 0.0180431694 0.0057637902 0.0015654739
[11] 0.0003652772
```

b) berechnen Sie die Wahrscheinlichkeitsverteilung des Poissonmodells $P(3)$ und vergleichen Sie es mit dem binomialen Modell $B(30, 0.1)$.

```
result <- data.frame(binomial= dbinom(c(0,1,2,3,4,5,6,7,8,9,10),
                                      size = 30, prob = 0.1),
                     poisson = dpois(c(0,1,2,3,4,5,6,7,8,9,10),
                                     lambda = 3))
head(result)
```

binomial	poisson
0.0423912	0.0497871
0.1413039	0.1493612
0.2276562	0.2240418
0.2360879	0.2240418
0.1770659	0.1680314
0.1023048	0.1008188

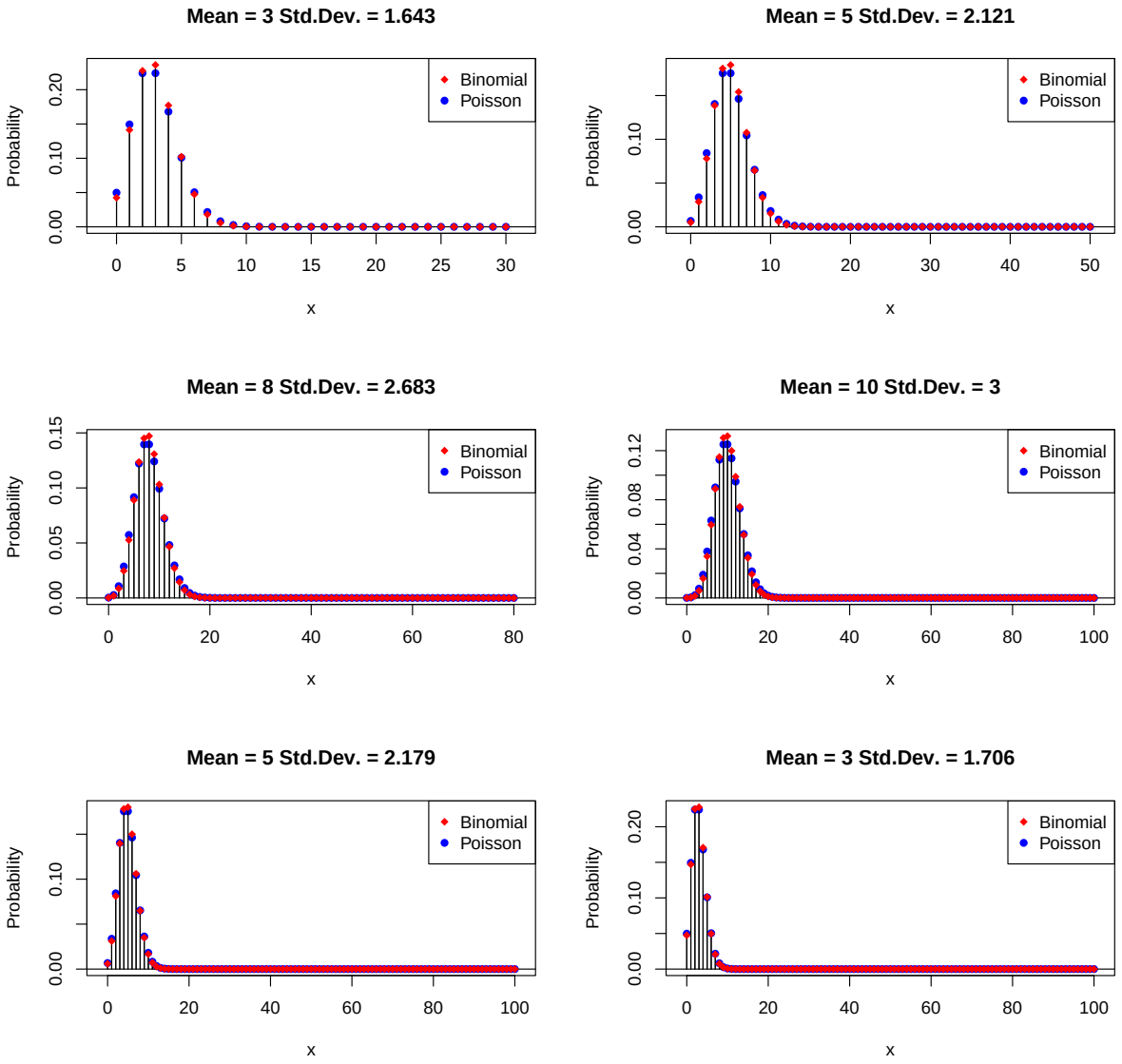
c) berechnen Sie die Wahrscheinlichkeitsverteilung des binomialen Modells $B(100, 0.3)$ und vergleichen Sie es es mit dem Modell $P(3)$. Sind diese Modelle ähnlicher als die vorherigen?

```
result <- data.frame(binomial= dbinom(c(0,1,2,3,4,5,6,7,8,9,10),
                                      size = 100, prob = 0.3),
                     poisson = dpois(c(0,1,2,3,4,5,6,7,8,9,10),
                                     lambda = 3))
head(result)
```

binomial	poisson
0	0.0497871
0	0.1493612
0	0.2240418
0	0.2240418
0	0.1680314
0	0.1008188

d) Plotten Sie die Wahrscheinlichkeitsfunktionen der vorherigen Modelle. Erhöhen Sie die Anzahl der Wiederholungen und verringern Sie die Erfolgswahrscheinlichkeit im Binomialmodell und beobachten Sie, wie sich die Wahrscheinlichkeiten des Binomialmodells und des Poissonmodells annähern.

```
# um nicht immer wieder den selben Plot-Befehl aufzurufen
# erstellen wir eine Hilfsfunktion
#-----------------------------------------------
myplot <- function(n, p){
  # vorberechnen
  mu <- p*n
  sd <- sqrt(n*p*(1-p))
  # plotten
  plot( seq(0,n), dpois( seq(0,n), mu ), type="h",
       xlim=c(-1,n+1), xlab="x", ylab="Probability",
       ylim=range(0,dpois( seq(0,n), mu), dbinom(seq(0,n),n,p)))
  points( seq(0,n), dpois( seq(0,n), mu ), pch=16, col="blue")
  points( seq(0,n), dbinom( seq(0,n), n, p), type="h")
  abline(h=0)
  points( seq(0,n), dbinom( seq(0,n), n, p), pch=18, col="red" )
  title( paste("Mean", "=", round(mu,3), "Std.Dev.", "=", round(sd,3)))
  legend("topright", c("Binomial", "Poisson"),
         col = c("red","blue"), pch = c(18,16))   }
#-----------------------------------------------

# plots vergleichen
myplot(30, 0.1)
myplot(50, 0.1)
myplot(80, 0.1)
myplot(100, 0.1)
myplot(100, 0.05)
myplot(100, 0.03)
```



2.38 Lösung zur Aufgabe 1.6.4

Wie groß ist die Wahrscheinlichkeit, beim Werfen von 100 Münzen zwischen 40 und 60 Mal `Kopf` zu erhalten (beide Werte eingeschlossen)?

```
sum(dbinom(40:60, size = 100, prob = 0.5))
```

```
[1] 0.9647998
```

2.39 Lösung zur Aufgabe 1.6.5

a) wie groß ist die Wahrscheinlichkeit, dass die Hälfte der Patienten geheilt wird?

```
# n=6 Patienten
# p =0.85
# k=3 (die Hälfte von 6)
dbinom(3, size = 6, prob = 0.85)
```

```
[1] 0.04145344
```

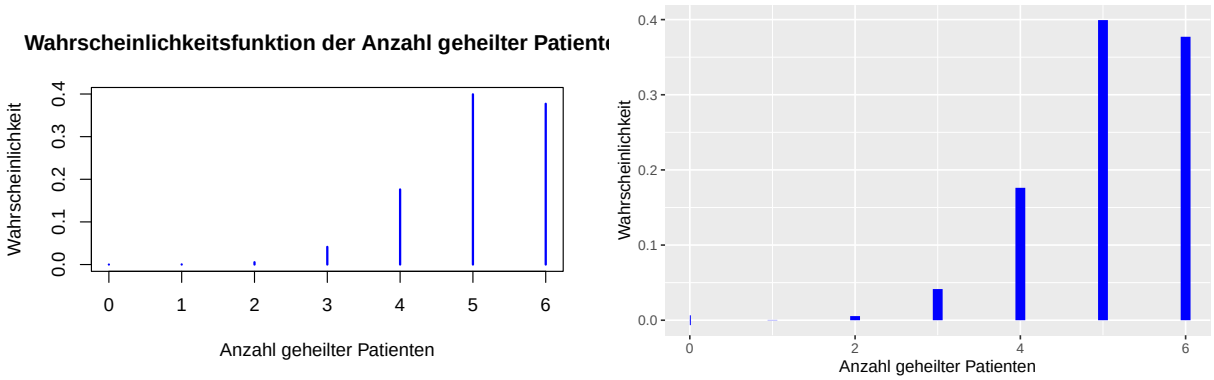
b) wie groß ist die Wahrscheinlichkeit, dass mindestens 4 Patienten geheilt werden?

```
pbinom(3, size = 6, prob = 0.85, lower.tail = FALSE)
```

```
[1] 0.9526614
```

c) plotten Sie die Wahrscheinlichkeitsfunktion für die Anzahl geheilter Patienten.

```
df <- data.frame(x=0:6,
                 y=dbinom(0:6, size = 6, prob = 0.85))

# plot()
plot(df$x, df$y, type="h", lwd=2, col="blue",
     xlab = "Anzahl geheilter Patienten",
     ylab = "Wahrscheinlichkeit",
     main = "Wahrscheinlichkeitsfunktion der Anzahl geheilter Patienten")
# ggplot
ggplot(df, aes(x=x, y=0, xend=x, yend=y)) +
  geom_segment(col="blue", lwd=3) +
  xlab("Anzahl geheilter Patienten") +
  ylab("Wahrscheinlichkeit")
```



2.40 Lösung zur Aufgabe 1.6.6

Die Wahrscheinlichkeit einer starken Impfreaktion beträgt 0,001. Wenn 2.000 Personen geimpft werden, wie hoch ist die Wahrscheinlichkeit für starke Reaktionen?

```
# n=2000 Patienten
# p =0.001
# k=1
pbinom(1, size = 2000, prob = 0.001, lower.tail=FALSE)
```

```
[1] 0.5941296
```

2.41 Lösung zur Aufgabe 1.6.7

a) Wie hoch ist die Wahrscheinlichkeit, dass weniger als 4 Anrufe in 2 Sekunden eintreffen?

```
# 120 Anrufe pro Minute sind
# 2 Anrufe pro Sekunde
# lambda für 2 Sekunden ist also 4
ppois(3, lambda=4, lower.tail=TRUE)
```

```
[1] 0.4334701
```

b) Wie hoch ist die Wahrscheinlichkeit, dass mindestens 3 Anrufe in 3 Sekunden eintreffen?

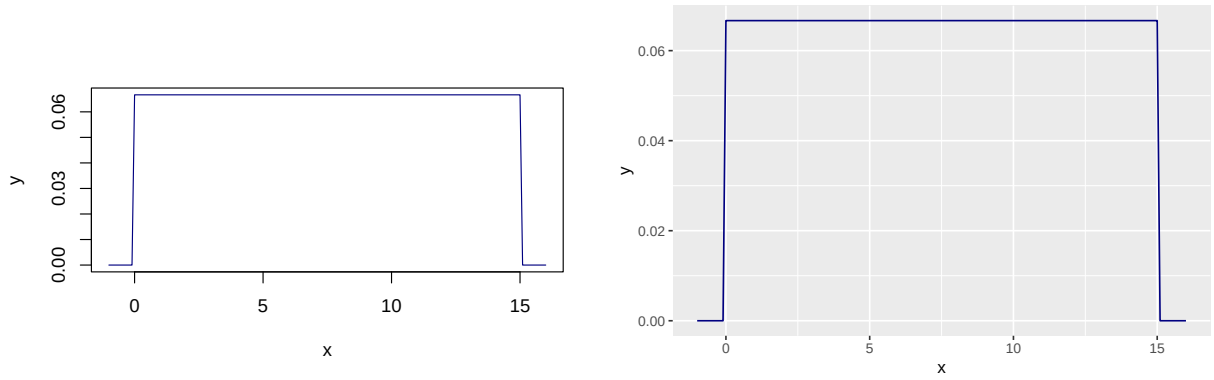
```
ppois(2, lambda=6, lower.tail=FALSE)
```

```
[1] 0.9380312
```

2.42 Lösung zur Aufgabe 1.7.1

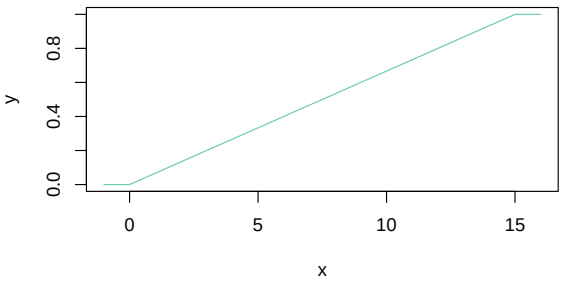
a) Plotten Sie die Dichtefunktion der Wartezeit.

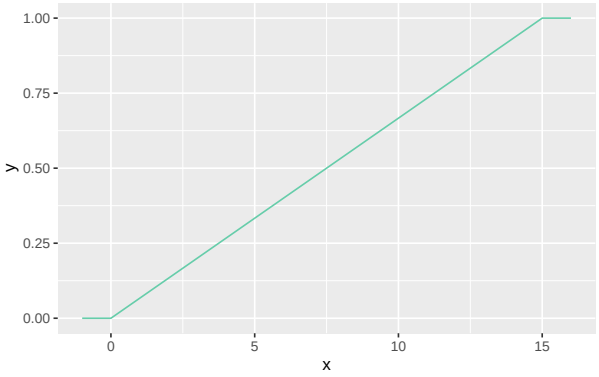
```
# x-Werte
x <- seq(-1, 16, by=0.1)
# Dichtefunktion der Uniformverteilung für alle x
y <- dunif(x, min=0, max=15)
# Datenframe
df <- data.frame(x, y)

# plot()
plot(x,y, type="l", col="navyblue")
# ggplot()
ggplot(df, aes(x=x, y=y)) +
  geom_line(col="navyblue")
```



b) Plotten Sie die Verteilungsfunktion der Wartezeit.

```
# x-Werte
x <- seq(-1, 16, by=0.1)
# Verteiluingsfunktion der Uniformverteilung für alle x
y <- punif(x, min=0, max=15)
# Datenframe
df <- data.frame(x, y)

# plot()
plot(x,y, type="l", col="aquamarine3")
# ggplot()
ggplot(df, aes(x=x, y=y)) +
  geom_line(col="aquamarine3")
```





c) Berechnen Sie die Wahrscheinlichkeit, weniger als 5 Minuten auf den Bus zu warten.

```
punif(5, min = 0, max = 15, lower.tail=TRUE)
```

```
[1] 0.3333333
```

d) Berechnen Sie die Wahrscheinlichkeit, länger als 12 Minuten auf den Bus zu warten.

```
punif(12, min = 0, max = 15, lower.tail=FALSE)
```

```
[1] 0.2
```

Die Wahrscheinlichkeit beträgt 20%.

e) Berechnen Sie die Wahrscheinlichkeit, zwischen 5 und 10 Minuten auf den Bus zu warten.

```
punif(10, min = 0, max = 15) - punif(5, min = 0, max = 15)
```

```
[1] 0.3333333
```

Die Wahrscheinlichkeit beträgt 33,33%.

f) Bei welcher Zeit zwischen 0 und 15 Minuten muss die Hälft der Personen kürzer auf den Bus warten als die angegebene Zeit?

```
qunif(0.5, min = 0, max = 15, lower.tail=TRUE)
```

```
[1] 7.5
```

50% der Personen muss weniger als 7,5 Minuten auf den Bus warten.

g) Bei welcher Zeit zwischen 0 und 15 Minuten müssen 10% der Personen länger auf den Bus warten als die angegebene Zeit?

```
qunif(0.1, min = 0, max = 15, lower.tail=FALSE)
```

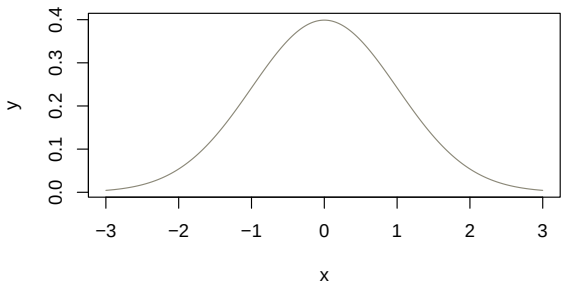
```
[1] 13.5
```

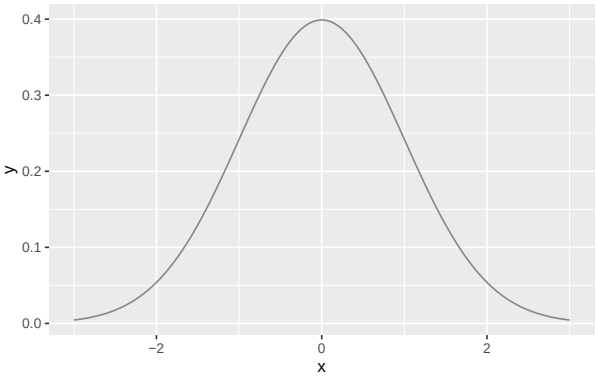
10% der Personen müssen länger als 13,5 Minuten auf den Bus warten.

2.43 Lösung zur Aufgabe 1.7.2

a) Plotten Sie die Dichtefunktion von Z.

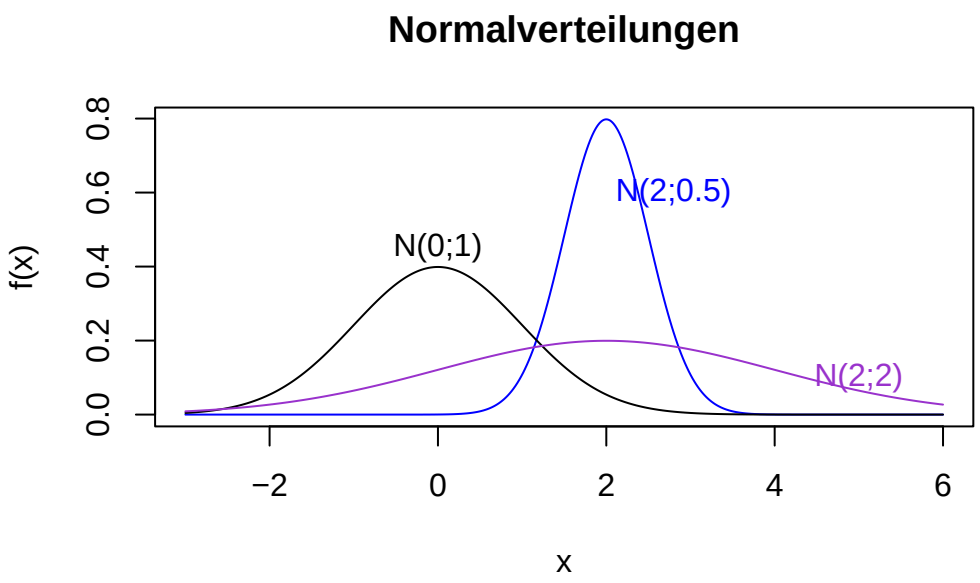
```
x <- seq(-3, 3, 0.01)
y <- dnorm(x, mean = 0, sd = 1)

df = data.frame(x = x, y = y)

# plot()
plot(x,y, type="l", col="cornsilk4")
# ggplot()
ggplot(df, aes(x=x, y=y)) +
  geom_line(col="cornsilk4")
```





b) Wie beeinflussen Mittelwert und Standardabweichung die Form der Gausschen Glockenkurve?

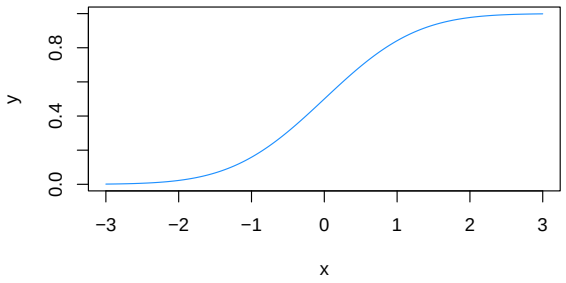
```
# erzeuge neue Werte von -3 bis 6
x <- seq(-3,6, by=0.005)

# Alles zusammen plotten
plot(x,dnorm(x,mean=2,s=0.5), col="blue", type="l", xlab="x",
     ylab="f(x)",main="Normalverteilungen")
lines(x,dnorm(x,mean=0,s=1), col="black")
lines(x,dnorm(x,mean=2,s=2), col="darkorchid")
text(0,.45,"N(0;1)")
text(2.8, 0.6, "N(2;0.5)", col="blue")
text(5, 0.1, "N(2;2)", col="darkorchid")
```

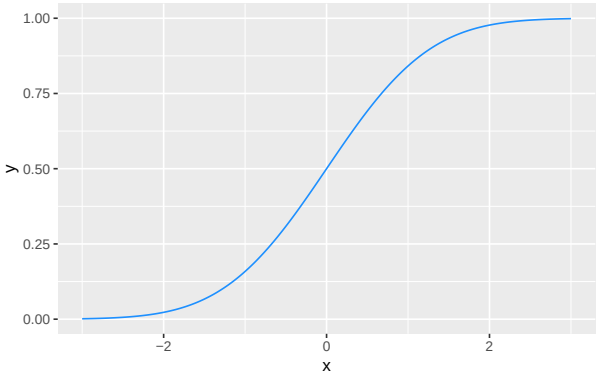


Der MIttelwert verschiebt die Kurve, die Standardabweichung verformt sie. Je größer die Standardabweichung, desto flacher und breiter ist die Kurve.

c) Plotten Sie die Verteilungsfunktion von Z.

```
# erzeuge neue Werte von -3 bis 3
x <- seq(-3, 3, 0.01)
y <- pnorm(x, mean=0, sd=1)
df = data.frame(x, y)

# plot()
plot(x,y, type="l", col="dodgerblue1")
# ggplot()
ggplot(df, aes(x=x, y=y)) +
  geom_line(col="dodgerblue1")
```





d) Berechnen Sie die Wahrscheinlichkeit $P(Z < -1)$.

```
pnorm(-1, mean=0, sd=1, lower.tail=TRUE)
```

```
[1] 0.1586553
```

e) Berechnen Sie die Wahrscheinlichkeit $P(Z > 1)$.

```
pnorm(1, mean=0, sd=1, lower.tail=FALSE)
```

```
[1] 0.1586553
```

f) Berechnen Sie die Wahrscheinlichkeit, dass Z zwischen dem Mittelwert minus der Standardabweichung und dem Mittelwert plus der Standardabweichung liegt, d. h. $P(-1 \leq Z \leq 1)$.

```
pnorm(1, mean=0, sd=1) - pnorm(-1, mean=0, sd=1)
```

```
[1] 0.6826895
```

g) Berechnen Sie die Wahrscheinlichkeit, dass Z zwischen dem Mittelwert minus zwei Standardabweichungen und dem Mittelwert plus zwei Standardabweichungen liegt, d. h. $P(-2 \leq Z \leq 2)$.

```
pnorm(2, mean=0, sd=1) - pnorm(-2, mean=0, sd=1)
```

```
[1] 0.9544997
```

h) Berechnen Sie die Wahrscheinlichkeit, dass Z zwischen dem Mittelwert minus drei Standardabweichungen und dem Mittelwert plus drei Standardabweichungen liegt, d. h. $P(-3 \leq Z \leq 3)$.

```
pnorm(3, mean=0, sd=1) - pnorm(-3, mean=0, sd=1)
```

```
[1] 0.9973002
```

i) Berechnen Sie die Quartile.

```
qnorm(c(0.25, 0.5, 0.75), mean=0, sd=1)
```

```
[1] -0.6744898  0.0000000  0.6744898
```

j) Bei welchem Z-Wert liegen 95% der Fläche unterhalb des Wertes?

```
qnorm(0.95, mean=0, sd=1, lower.tail=TRUE)
```

```
[1] 1.644854
```

k) Bei welchem Z-Wert liegen 2,5% der Fläche oberhalb des Wertes?

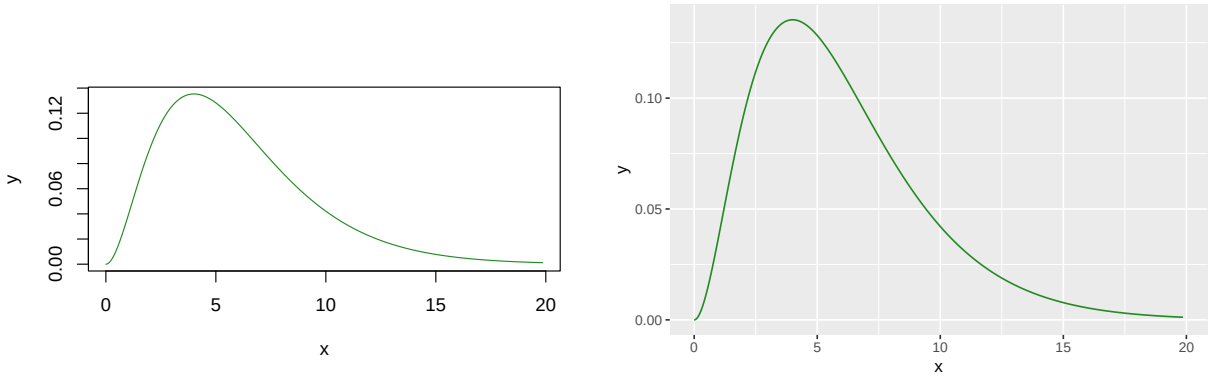
```
qnorm(0.025, mean=0, sd=1, lower.tail=FALSE)
```

```
[1] 1.959964
```

2.44 Lösung zur Aufgabe 1.7.3

a) Plotten Sie die Dichtefunktion dieser Verteilung.

```
x <- seq(0, 19.86, 0.01)
y <- dchisq(x, df=6)

df = data.frame(x = x, y = y)

# plot()
plot(x,y, type="l", col="forestgreen")
# ggplot()
ggplot(df, aes(x=x, y=y)) +
  geom_line(col="forestgreen")
```



b) Wie groß ist die Wahrscheinlichkeit für $P(X < 6)$?

```
pchisq(6, df=6, lower.tail=TRUE)
```

```
[1] 0.5768099
```

c) Berechnen Sie das fünfte Perzentil der Verteilung.

```
qchisq(0.05, df=6)
```

```
[1] 1.635383
```

d) Bei welchem Wert liegen 10% der Fläche oberhalb des Wertes?

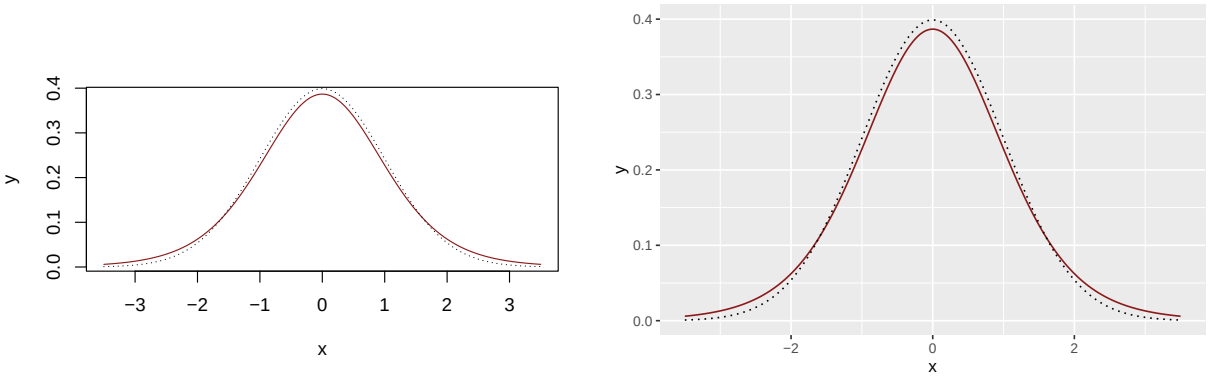
```
qchisq(0.1, df=6, lower.tail=FALSE)
```

```
[1] 10.64464
```

2.45 Lösung zur Aufgabe 1.7.4

a) Plotten Sie die Dichtefunktion von X und vergleichen Sie diese mit der Dichtefunktion der Standardnormalverteilung.

```
x <- seq(-3.5, 3.5, 0.01)
y <- dt(x, df=8)
y2 <- dnorm(x)
df = data.frame(x=x, y=y, y2=y2)

# plot()
plot(x,y, type="l", col="firebrick4")
lines(x,y2, lty=3)
# ggplot()
ggplot(df, aes(x=x, y=y)) +
  geom_line(col="firebrick4") +
  geom_line(aes(x=x, y=y2), lty=3)
```



b) Berechnen Sie das 8te Perzentil von X.

```
qt(0.08, df=8)
```

```
[1] -1.548892
```

c) Bei welchem Wert von X liegen 5% aller Fälle oberhalb dieses Wertes?

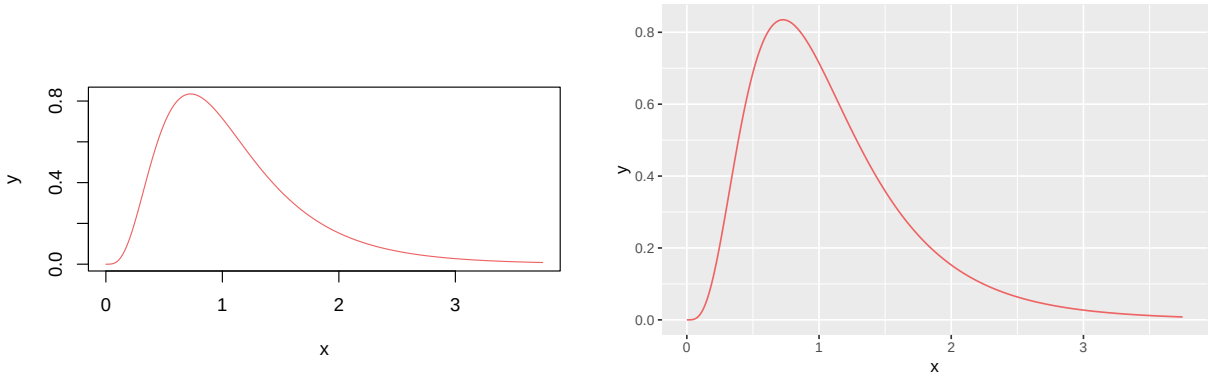
```
qt(0.05, df=8, lower.tail = FALSE)
```

```
[1] 1.859548
```

2.46 Lösung zur Aufgabe 1.7.5

a) Plotten Sie die Dichtefunktion von X.

```
x <- seq(0, 3.75, 0.01)
y <- df(x, df1 = 10, df2 = 20)

df = data.frame(x=x, y=y)

# plot()
plot(x,y, type="l", col="indianred2")
# ggplot()
ggplot(df, aes(x=x, y=y)) +
  geom_line(col="indianred2")
```



b) Berechnen Sie Wahrscheinlichkeit $P(X > 1)$.

```
pf(1, df1=10, df2=20, lower.tail=FALSE)
```

```
[1] 0.4755005
```

c) Berechnen Sie den Interquartilsabstand.

```
qf(c(0.75), df1=10, df2=20)  - qf(0.25, df1=10, df2=20)
```

```
[1] 0.7430938
```

2.47 Lösung zur Aufgabe 1.7.6

a) Berechnen Sie die Wahrscheinlichkeit, dass ein zufällig ausgewählter Diabetiker einen Glukosespiegel von weniger als 120 mg/100 ml hat.

```
pnorm(120, mean=106, sd=8)
```

```
[1] 0.9599408
```

b) Wie viel Prozent der Personen haben einen Glukosespiegel zwischen 90 und 120 mg/100 ml?

```
pnorm(120, mean=106, sd=8) - pnorm(90, mean=106, sd=8)
```

```
[1] 0.9371907
```

Etwa 93.72% der Personen.

c) Berechnen und interpretieren Sie das erste Quartil des Glukosespiegels.

```
qnorm(0.25, mean=106, sd=8)
```

```
[1] 100.6041
```

2.48 Lösung zur Aufgabe 1.7.7

a) Wie viele von ihnen haben einen Cholesterinspiegel zwischen 210 und 240 mg/dl?

```
# Anteile berechnen
pnorm(240, mean=220, sd=30) - pnorm(210, mean=220, sd=30)
```

```
[1] 0.3780661
```

Etwa 37.81% der Personen.

b) Wenn ein Cholesterinspiegel von mehr als 250 mg/dl eine Thrombose auslösen kann, wie viele von ihnen sind thrombosegefährdet?

```
pnorm(250, mean=220, sd=30, lower.tail=FALSE)
```

```
[1] 0.1586553
```

Etwa 15.87% der Personen.

c) Welcher Cholesterinwert wird von mindestens 20% der Männer erreicht?

```
# Anteile berechnen
qnorm(0.2, mean=220, sd=30)
```

```
[1] 194.7514
```

2.49 Lösung zur Aufgabe 1.8.1

a) Übertragen Sie die Daten in ein Datenframe mit der Variable `Konzentration`.

```
Konzentration <- c(17.6, 19.2, 21.3, 15.1, 17.6, 18.9, 16.2, 18.3, 19.0, 16.4)
```

b) Berechnen Sie das Konfidenzintervall für die mittlere Konzentration bei einem Konfidenzniveau von 95% (Signifikanzlevel $\alpha = 0,05$).

```
d <-t.test(Konzentration, mu=0, conf.level=0.95)
d
```

```
    One Sample t-test

data:  Konzentration
t = 31.78, df = 9, p-value = 1.485e-10
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 16.68158 19.23842
sample estimates:
mean of x
    17.96
```

```
# nur Konfidenzintervall ausgeben
as.numeric(d$conf.int)
```

```
[1] 16.68158 19.23842
```

c) Berechnen Sie das Konfidenzintervall für die mittlere Konzentration bei einem Konfidenzniveau von 99% (Signifikanzlevel $\alpha = 0,01$).

```
d <-t.test(Konzentration, mu=0, conf.level=0.99)
d
```

```
    One Sample t-test

data:  Konzentration
t = 31.78, df = 9, p-value = 1.485e-10
```

```
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 16.1234 19.7966
sample estimates:
mean of x
    17.96
```

```
# nur Konfidenzintervall ausgeben
as.numeric(d$conf.int)
```

```
[1] 16.1234 19.7966
```

d) Wenn wir die Genauigkeit des Intervalls als den Kehrwert seiner Breite definieren, wie ändert sich die Genauigkeit eines Intervalls, wenn wir das Konfidenzniveau erhöhen?

Mit höherem Konfidenzniveau sinkt die Genauigkeit der Aussagen.

e) Welche Stichprobengröße wird benötigt, um den mittleren Konzentrationswert mit einem Fehler von $\pm 0.5 \text{mg/mm}^3$ und einem Konfidenzniveau von 95% Sicherheit zu bestimmen?

```
# Berechnung der Standardabweichung der Stichprobe
sigma <- sd(Konzentration)

# Gegebene Werte
# z-Wert für 95% Konfidenzniveau
z <- 1.96

# Fehlermarge
E <- 0.5

# Berechnung der Stichprobengröße
n <- (z * sigma / E)^2

# Aufrunden auf die nächste ganze Zahl
ceiling(n)
```

```
[1] 50
```

f) Wenn die Konzentration des Wirkstoffs mindestens 16 mg/mm^3 betragen muss, um wirksam zu sein, ist dann unsere Medikamentencharge wirksam?

```
t.test(Konzentration, mu = 16, alternative = "greater")
```

```
    One Sample t-test

data:  Konzentration
t = 3.4682, df = 9, p-value = 0.003534
```

```
alternative hypothesis: true mean is greater than 16
95 percent confidence interval:
 16.92404      Inf
sample estimates:
mean of x
    17.96
```

Der Test ist signifikant. Wir können also sagen, dass unserer Medikamentencharge wirksam ist.

2.50 Lösung zur Aufgabe 1.8.2

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `Hof` und `Fett`.

```
# Daten übertragen
Hof1 <- data.frame(Fett = c(0.34, 0.34,
          0.32, 0.35,
          0.33, 0.33,
          0.32, 0.32,
          0.33, 0.30,
          0.31, 0.32))
Hof1$Hof <- "Hof 1"

Hof2 <- data.frame(Fett = c(0.28, 0.29,
          0.30, 0.32,
          0.32, 0.31,
          0.29, 0.29,
          0.31, 0.32,
          0.29, 0.31,
          0.33, 0.32,
          0.32, 0.33))
Hof2$Hof <- "Hof 2"

milch <- rbind(Hof1, Hof2)
milch$Hof <- factor(milch$Hof)
```

b) Berechnen Sie das 95%-Konfidenzintervall für den durchschnittlichen Fettgehalt.

```
t.test(milch$Fett, conf.level=0.95)
```

```
    One Sample t-test

data:  milch$Fett
t = 96.537, df = 27, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.3090040 0.3224246
```

```
sample estimates:
mean of x
0.3157143
```

c) Berechnen Sie das 95%-Konfidenzintervall für den durchschnittlichen Fettgehalt, getrennt nach Höfen.

```r
t.test(Hof1$Fett, conf.level=0.95)
```

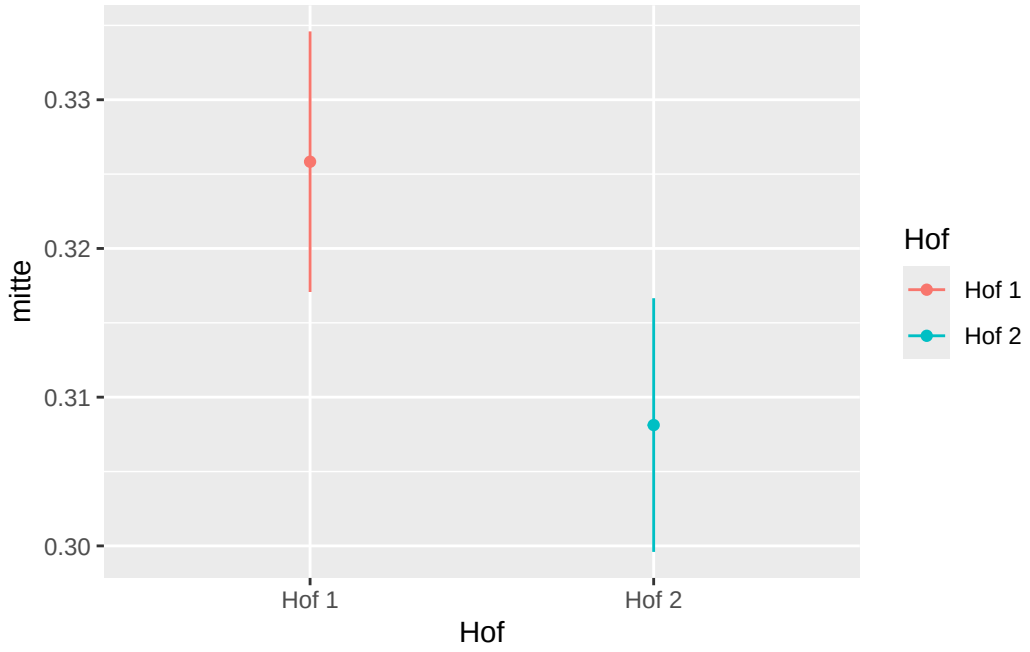
```

    One Sample t-test

data:  Hof1$Fett
t = 81.853, df = 11, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.3170719 0.3345948
sample estimates:
mean of x
0.3258333
```

```r
t.test(Hof2$Fett, conf.level=0.95)
```

```

    One Sample t-test

data:  Hof2$Fett
t = 76.994, df = 15, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.299595 0.316655
sample estimates:
mean of x
 0.308125
```

d) Plotten Sie das 95%-Konfidenzintervall für den durchschnittlichen Fettgehalt, getrennt nach Höfen..

```
# Vorbereitung
h1 <- t.test(Hof1$Fett, conf.level=0.95)
h2 <- t.test(Hof2$Fett, conf.level=0.95)

# Als Datenframe für ggplot zusammenbauen
df <- data.frame(unten = c(h1$conf.int[1], h2$conf.int[1]),
                 oben  = c(h1$conf.int[2], h2$conf.int[2]),
                 mitte = c(h1$estimate, h2$estimate ),
                 Hof = c("Hof 1", "Hof 2"))

# ggplot
ggplot(df, aes(x=Hof, color=Hof)) +
  geom_point(aes(y=mitte)) +
  geom_segment(aes(xend=Hof, y=unten, yend=oben))
```



e) Lässt sich aus den Konfidenzintervallen ein signifikanter Untschied zwischen den Höfen feststellen?

```
df
```

unten	oben	mitte	Hof
0.3170719	0.3345948	0.3258333	Hof 1
0.2995950	0.3166550	0.3081250	Hof 2

Die beiden Konfidenzintervalle überschneiden sich nicht. Das heisst, es kann ein signifikanter Unterschied abgeleitet werden.

2.51 Lösung zur Aufgabe 1.8.3

a) Übertragen Sie die Daten in ein Datenframe mit der Variable `Antwort`.

```
# Daten übertragen
Antwort <- c("nein",  "ja",  "nein",  "nein",  "nein",  "ja",  "nein",
             "ja", "ja",  "ja",  "ja",  "nein",  "ja",  "nein", "ja",
             "nein",  "nein", "nein",  "ja",  "ja",  "ja",  "nein",
             "nein",  "ja",  "nein",  "nein", "ja",  "ja",  "nein",
             "nein",  "ja",  "nein",  "ja", "nein")
```

b) Berechnen Sie das Konfidenzintervall für den Anteil an Studierenden, welche die Bibliothek wöchentlich nutzen mit einem Signifikanzlevel von $\alpha = 0,01$.

```
freq <- table(Antwort)
bib <- prop.test(freq[["ja"]], sum(freq),
                 alternative="two.sided",
                 p=0.5, conf.level=0.99)
bib
```

```
    1-sample proportions test with continuity correction

data:  freq[["ja"]] out of sum(freq), null probability 0.5
X-squared = 0.029412, df = 1, p-value = 0.8638
alternative hypothesis: true p is not equal to 0.5
99 percent confidence interval:
 0.2617050 0.6896622
sample estimates:
        p
0.4705882
```

c) Wie präzise ist das Intervall?

```
bib$conf.int[2] - bib$conf.int[1]
```

```
[1] 0.4279572
```

Das Intervall ist sehr breit und daher unpräzise.

d) Welcher Stichprobenumfang ist erforderlich, um eine Schätzung des Anteils der Studenten zu erhalten, die die Bibliothek mindestens einmal pro Woche nutzen, mit einem Fehler von ±1% und einem Konfidenzniveau von 95%?

```
# gemessene Proportionen
prop <- bib$estimate

# Z-Wert für 95% Konfidenz
z <- 1.96

# Fehlerspanne +-1%
e <- 0.01

# Fallzahl berechnen
n <- (z^2 * prop * (1 - prop)) / (e^2)
```

Es werden 9571 Probanden benötigt.

2.52 Lösung zur Aufgabe 1.8.4

a) Berechnen Sie das 95%-Konfidenzintervall für den Anteil an geimpften Probanden in der Grundgesamtheit.

```
# Daten übertragen
prop.test(154, 200, p=0.5, conf.level=0.95)
```

```
    1-sample proportions test with continuity correction

data:  154 out of 200, null probability 0.5
X-squared = 57.245, df = 1, p-value = 3.848e-14
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.7042503 0.8251428
sample estimates:
   p
0.77
```

b) Wenn das Gesundheitsministerium das Ziel verfolgt, dass mindestens 70% der Menschen über 65 mit Atemwegserkrankungen geimpft sind, können wir dann sagen, dass das Ministerium das Ziel erreicht hat?

```
# Daten übertragen
prop.test(154, 200, p=0.5, conf.level=0.95)$conf.int[1]
```

```
[1] 0.7042503
```

Da die untere Grenze des Konfidenzintervalls größer als 0,7 ist, können wir bestätigen, dass das

Ministerium sein Ziel erreicht hat.

2.53 Lösung zur Aufgabe 1.8.5

a) Berechnen Sie die Konfidenzintervalle für den Mittelwert mit den Signifikanzniveaus 0.1, 0.05 und 0.01.

```
# Daten übertragen
cholesterin <- c(196, 212, 188, 206, 203, 210, 201, 198)

t.test(cholesterin, conf.level=0.90)
```

```
    One Sample t-test

data:  cholesterin
t = 72.849, df = 7, p-value = 2.416e-11
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 196.5031 206.9969
sample estimates:
mean of x
   201.75
```

```
t.test(cholesterin, conf.level=0.95)
```

```
    One Sample t-test

data:  cholesterin
t = 72.849, df = 7, p-value = 2.416e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 195.2014 208.2986
sample estimates:
mean of x
   201.75
```

```
t.test(cholesterin, conf.level=0.99)
```

```
    One Sample t-test

data:  cholesterin
t = 72.849, df = 7, p-value = 2.416e-11
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
```

```
 192.0585 211.4415
sample estimates:
mean of x
   201.75
```

b) Kann man schließen, dass der Mittelwert des Cholesterinspiegels der Bevölkerung unter 210 mg/dl liegt?

```
# Daten übertragen
as.numeric(t.test(cholesterin, conf.level=0.90)$conf.int)
```

```
[1] 196.5031 206.9969
```

```
as.numeric(t.test(cholesterin, conf.level=0.95)$conf.int)
```

```
[1] 195.2014 208.2986
```

```
as.numeric(t.test(cholesterin, conf.level=0.99)$conf.int)
```

```
[1] 192.0585 211.4415
```

Die obere Grenze des 99%-Konfidenzintervall ist größer als 210. Wird dies berücksichtigt, lässt dich die Aussage nicht bestätigen.

2.54 Lösung zur Aufgabe 1.8.6

a) Berechnen Sie für jede Therapie das 95% Konfidenzintervall für den Anteil an Personen, die geheilt wurden.

```
# Daten übertragen
a <- c(rep("geheilt", 18), rep("nicht", 7))
b <- c(rep("geheilt", 21), rep("nicht", 14))

freqA <- table(a)
freqB <- table(b)
A <- prop.test(freqA[["geheilt"]], sum(freqA),
                 alternative="two.sided",
                 p=0.5, conf.level=0.95)
B <- prop.test(freqB[["geheilt"]], sum(freqB),
                 alternative="two.sided",
                 p=0.5, conf.level=0.95)
# Konfidenzintervalle
A$conf.int[2] - A$conf.int[1]
```

```
[1] 0.3672662
```

```
B$conf.int[2] - B$conf.int[1]
```

```
[1] 0.3343891
```

Das Konfidenzintervall von Gruppe B ist schmaler, und damit auch präziser.

2.55 Lösung zur Aufgabe 1.8.7

```
# lade Datensatz
load(url("https://www.produnis.de/R/data/neonates.RData"))
```

a) Berechnen Sie das 99% Konfidenzintervall für den Mittelwert des Gewichts der Neugeborenen.

```
# t-Test
d <- t.test(neonates$weight, conf.level = 0.99)

# Konfidenzgrenzen
as.numeric(d$conf.int)
```

```
[1] 2.975844 3.075531
```

b) Berechnen Sie die Konfidenzintervalle für den APGAR-Score nach 1 Minute und für den APGAR-Score nach 5 Minuten und vergleiche sie beide Intervalle. Gibt es auf Grundlage der Konfidenzintervalle einen signifikanten Unterschied zwischen den Mittelwerten der beiden Scores?

```
# t-Test
a1 <- t.test(neonates$apgar1, conf.level = 0.99)
a5 <- t.test(neonates$apgar5, conf.level = 0.99)

# Konfidenzgrenzen
as.numeric(a1$conf.int)
```

```
[1] 5.422182 5.834068
```

```
as.numeric(a5$conf.int)
```

```
[1] 5.998597 6.426403
```

Die Konfidenzgrenzen schneiden sich nicht. Das bedeutet, wir können von einem signifikanten Unterschied ausgehen.

c) Berechnen Sie die Konfidenzintervalle für den Prozentsatz der Neugeborenen mit einem Gewicht von $\leq 2,5$ kg für Raucher- und Nichtrauchermütter und vergleichen Sie die Intervalle.

```
# geringes Gewicht kategorisieren
neonates$GG <- "normal"
neonates$GG[neonates$weight<2.50001] <- "low"

# Subsets bilden
k1 <- subset(neonates, smoke=="Yes")
k2 <- subset(neonates, smoke=="No")
# Häufigkeitstabelle
freq1 <- table(k1$GG)
freq2 <- table(k2$GG)

# Proportion Test
m1 <- prop.test(freq1[["low"]], sum(freq1),
                alternative="two.sided",
                p=0.5, conf.level=0.95)

m2 <- prop.test(freq2[["low"]], sum(freq2),
                alternative="two.sided",
                p=0.5, conf.level=0.95)

# Konfidenzgrenzen anzeigen
as.numeric(m1$conf.int)
```

```
[1] 0.1049334 0.2610870
```

```
as.numeric(m2$conf.int)
```

```
[1] 0.008396857 0.055169043
```

Die Konfidenzgrenzen schneiden sich nicht. Das bedeutet, wir können von einem signifikanten Unterschied ausgehen.

2.56 Lösung zur Aufgabe 1.9.1

a) Erstellen Sie ein Datenframe mit den Variablen `vorher` und `nachher` und übertragen Sie die Daten.

```
df <- data.frame(
  vorher = c(147, 163, 121, 205, 132, 190, 176, 147),
  nachher = c(150, 171, 132, 208, 141, 184, 182, 145)
)
```

b) Berechnen Sie den Mittelwert der monatlichen Umsätze vor und nach der Kampagne. Sind die Mittelwerte unterschiedlich? Hat die Kampagne den Absatz des Arzneimittels erhöht?

```
mean(df$vorher)
```

```
[1] 160.125
```

```
mean(df$nachher)
```

```
[1] 164.125
```

Der Mittelwert ist nach der Kampagne höher.

c) Berechnen Sie die Konfidenzintervalle für den durchschnittlichen Unterschied mit $\alpha = 0,05$ und $\alpha = 0,01$.

```
# t-Tests durchführen
a5 <- t.test (df$vorher, df$nachher, paired=TRUE, conf.level=0.95)
a1 <- t.test (df$vorher, df$nachher, paired=TRUE, conf.level=0.99)

# nur Konfidenzintervalle anzeigen
as.numeric(a5$conf.int)
```

```
[1] -8.8129585  0.8129585
```

```
as.numeric(a1$conf.int)
```

```
[1] -11.122852   3.122852
```

Beide Intervalle "reißen" die 0. Das bedeutet, dass der wahre Unterschied auch `0` sein könnte.

d) Können wir dieselbe Schlussfolgerung ziehen, wenn wir die Verkäufe nach der Kampagne der beiden letzten Apotheken ändern und 190 statt 182 und 165 statt 145 angeben? Was passiert mit den Konfidenzintervallen?

```
# Daten neu eingeben
df <- data.frame(
  vorher = c(147, 163, 121, 205, 132, 190, 176, 147),
  nachher = c(150, 171, 132, 208, 141, 184, 190, 165)
)
# t-Tests durchführen
a5 <- t.test (df$vorher, df$nachher, paired=TRUE, conf.level=0.95)
a1 <- t.test (df$vorher, df$nachher, paired=TRUE, conf.level=0.99)

# nur Konfidenzintervalle anzeigen
as.numeric(a5$conf.int)
```

```
[1] -13.740228  -1.259772
```

```
as.numeric(a1$conf.int)
```

```
[1] -16.735113   1.735113
```

Das 95%-Intervall lässt einen Unterschied vermuten. Das 99%-Intervall hingegen enthält immer noch die 0.

2.57 Lösung zur Aufgabe 1.9.2

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `Hof1` und `Hof2`.

```
# Daten übertragen
Hof1 <- data.frame(Fett = c(0.34, 0.34,
          0.32, 0.35,
          0.33, 0.33,
          0.32, 0.32,
          0.33, 0.30,
          0.31, 0.32))
Hof1$Hof <- "Hof 1"

Hof2 <- data.frame(Fett = c(0.28, 0.29,
          0.30, 0.32,
          0.32, 0.31,
          0.29, 0.29,
          0.31, 0.32,
          0.29, 0.31,
          0.33, 0.32,
          0.32, 0.33))
Hof2$Hof <- "Hof 2"

milch <- rbind(Hof1, Hof2)
milch$Hof <- factor(milch$Hof)
```

b) Berechnen Sie das 95%-Konfidenzintervall für den durchschnittlichen Fettunterschied in der Milch von `Hof1` und `Hof2`.

```
# Daten übertragen
# zuvor auf gleiche Varainz prüfen
var.test(Hof1$Fett, Hof2$Fett)$p.value
```

```
[1] 0.626044
```

Der Test ist nicht signifikant, die Varianz ist in beiden Höfen gleich.

```r
# t.Test
d <- t.test(Fett ~ Hof, data=milch, var.equal=TRUE)
# Konfidenzintervall anzeigen
as.numeric(d$conf.int)
```

```
[1] 0.00584816 0.02956851
```

c) Kann man daraus schließen, dass der Unterschied zwischen den Milchfettmittelwerten der Betriebe signifikant ist? Welcher Betrieb hat Milch mit mehr Fett? Wie viel mehr Fett hat die Milch von `Hof1` als die Milch von `Hof2`?

```r
d <- t.test(Fett ~ Hof, data=milch, var.equal=TRUE)
d
```

```
    Two Sample t-test

data:  Fett by Hof
t = 3.0691, df = 26, p-value = 0.004973
alternative hypothesis: true difference in means between group Hof 1 and group Hof 2 is not
95 percent confidence interval:
 0.00584816 0.02956851
sample estimates:
mean in group Hof 1 mean in group Hof 2
          0.3258333           0.3081250
```

```r
# Unterschied
d$estimate[1] - d$estimate[2]
```

```
mean in group Hof 1
         0.01770833
```

Hof1 hat 0.0177083 mehr Fett in der Milch als Hof2. Da das Intervall die 0 nicht mit einschließt, können wir von einem Signifikanten Unterschied ausgehen.

2.58 Lösung zur Aufgabe 1.9.3

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `Antwort` und `Geschlecht`.

```
# Daten übertragen
df <- data.frame(
  Antwort = c("nein", "ja", "nein", "nein", "nein", "ja", "nein",
              "ja", "ja", "ja", "ja", "nein", "ja", "nein", "ja",
              "nein", "nein",  "nein", "ja", "ja", "ja", "nein",
              "nein", "ja", "nein", "nein", "ja", "ja", "nein",
              "nein", "ja", "nein", "ja", "nein"),
  Geschlecht = c("m", "w", "w", "m", "m", "m", "w", "w", "w", "w",
                 "m", "m", "w", "m", "w", "m", "m", "w", "m", "w",
                 "w", "w", "m", "w", "m", "m", "w", "w", "m", "m",
                 "w", "w", "w", "m")
)
```

b) Berechnen Sie das Konfidenzintervall für den Unterschied zwischen den Anteilen der Frauen und Männern, die die Bibliothek mindestens einmal pro Woche nutzen.

```
freq <- table(df)
prop.test(c(freq[["ja","m"]], freq[["ja","w"]]),
          c(sum(freq[,"m"]), sum(freq[,"w"])),
          alternative="two.sided", conf.level=0.95)
```

```
    2-sample test for equality of proportions with continuity correction

data:  c(freq[["ja", "m"]], freq[["ja", "w"]]) out of c(sum(freq[, "m"]), sum(freq[, "w"]))
X-squared = 7.6937, df = 1, p-value = 0.005541
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.8755141 -0.1939304
sample estimates:
   prop 1    prop 2
0.1875000 0.7222222
```

Das Konfidenzintervall schließt die 0 nicht mit ein. Wir können von einem signifikanten Unterschied ausgehen.

2.59 Lösung zur Aufgabe 1.9.4

```
df <- data.frame(course = c(rep("bestanden", 55), rep("durchgefallen", 25),
                            rep("bestanden", 32), rep("durchgefallen", 58)),
                 time = c(rep("morgens", 80), rep("abends", 90))

                 )
```

Gibt es signifikante Unterschiede zwischen den Prozentsätzen der Studierenden, die am Vormittag und am Nachmittag bestanden haben? Kann man daraus schließen, dass der Stundenplan die Ursache für diese Unterschiede ist?

```
freq <- table(df)
prop.test(c(freq[["bestanden","morgens"]], freq[["bestanden","abends"]]),
          c(sum(freq[,"morgens"]), sum(freq[,"abends"])),
          alternative="two.sided", conf.level=0.95)
```

```
    2-sample test for equality of proportions with continuity correction

data:  c(freq[["bestanden", "morgens"]], freq[["bestanden", "abends"]]) out of c(sum(freq[,
X-squared = 17.372, df = 1, p-value = 3.072e-05
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1783764 0.4855125
sample estimates:
   prop 1    prop 2
0.6875000 0.3555556
```

Das Konfidenzintervall enthält nicht die 0. Wir können also von einem signifikanten Unterschied ausgehen.

2.60 Lösung zur Aufgabe 1.9.5

```
df <- data.frame(vorher = c(182, 232, 191, 200, 148, 249, 276, 213, 241, 280, 262),
                 nachher = c(198, 210, 194, 220, 138, 220, 219, 161, 210, 213, 226)
                 )
```

a) Berechnen Sie das 95%-Konfidenzintervall für den durchschnittlichen Unterschied der Cholesterinwerte vor und nach den körperlichen Übungen.

```
t.test(df$vorher, df$nachher, paired=TRUE, conf.level=0.95)
```

```
    Paired t-test

data:  df$vorher and df$nachher
t = 2.756, df = 10, p-value = 0.02027
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
  4.614342 43.567477
sample estimates:
mean difference
       24.09091
```

b) Berechnen Sie das 99%-Konfidenzintervall für den durchschnittlichen Unterschied der Cholesterinwerte vor und nach den körperlichen Übungen

```
t.test(df$vorher, df$nachher, paired=TRUE, conf.level=0.99)
```

```
    Paired t-test

data:  df$vorher and df$nachher
t = 2.756, df = 10, p-value = 0.02027
alternative hypothesis: true mean difference is not equal to 0
99 percent confidence interval:
 -3.61228 51.79410
sample estimates:
mean difference
       24.09091
```

c) Auf Grundlage der zuvor berechneten Intervalle, welchen Schluss bezüglich des Einflusses von körperlichen Aktivitäten auf den Cholesterinspiegel können Sie ziehen?

```
d <- t.test(df$vorher, df$nachher, paired=TRUE, conf.level=0.99)
as.numeric(d$conf.int)
```

```
[1] -3.61228 51.79410
```

Das 99% Intervall enthält die 0. Auf diesem Niveau können wir den Unterschied nicht bestätigen.

2.61 Lösung zur Aufgabe 1.9.6

```
df <- data.frame(satisf = c(rep("zufrieden", 140), rep("unzufrieden", 60),
                            rep("zufrieden", 180), rep("unzufrieden", 120)),
                 haus = c(rep("Haus 1", 200), rep("Haus 2", 300))
                 )
```

a) Berechnen Sie das 95%-Konfidenzintervall für den Anteilsunterschied an zufriedenen Patienten in beiden Häusern.

```
freq <- table(df)
prop.test(c(freq[["zufrieden","Haus 1"]], freq[["zufrieden","Haus 2"]]),
          c(sum(freq[,"Haus 1"]), sum(freq[,"Haus 2"])),
          alternative="two.sided", conf.level=0.95)
```

```
    2-sample test for equality of proportions with continuity correction

data:  c(freq[["zufrieden", "Haus 1"]], freq[["zufrieden", "Haus 2"]]) out of c(sum(freq[, "
X-squared = 4.7833, df = 1, p-value = 0.02874
alternative hypothesis: two.sided
```

```
95 percent confidence interval:
 0.01153209 0.18846791
sample estimates:
prop 1 prop 2
   0.7    0.6
```

b) Wenn $\alpha = 0,01$ ist, können dann Rückschlüsse gezogen werden, ob der Unterschied der Anteile zufriedener Patienten signifikant ist?

```
prop.test(c(freq[["zufrieden","Haus 1"]], freq[["zufrieden","Haus 2"]]),
          c(sum(freq[,"Haus 1"]), sum(freq[,"Haus 2"])),
          alternative="two.sided", conf.level=0.99)
```

```
    2-sample test for equality of proportions with continuity correction

data:  c(freq[["zufrieden", "Haus 1"]], freq[["zufrieden", "Haus 2"]]) out of c(sum(freq[, "
X-squared = 4.7833, df = 1, p-value = 0.02874
alternative hypothesis: two.sided
99 percent confidence interval:
 -0.01495727  0.21495727
sample estimates:
prop 1 prop 2
   0.7    0.6
```

Wenn $\alpha = 0,01$ verwendet wird, enthält das Intervall die 0. Auf diesem Niveau können wir keinen signifikanten Unterschied zeigen.

2.62 Lösung zur Aufgabe 1.9.7

```
# lade Datensatz
load(url("https://www.produnis.de/R/data/neonates.RData"))
```

a) Berechnen Sie das 95%-Konfidenzintervall für den durchscnnittlichen Unterschied des Geburtsgewichts zwischen Kindern von Raucherinnen und Nichtraucherinnen. Wie groß ist der durchschnittliche Gewichtsunterschied?

```
# subsets
k1 <- subset(neonates, smoke=="Yes")
k2 <- subset(neonates, smoke=="No")

# t-Test
d <- t.test(k1$weight, k2$weight, conf.level = 0.95)

# Konfidenzgrenzen
as.numeric(d$conf.int)
```

```
[1] -0.3714943 -0.2179094
```

```
# Unterschied
d$estimate[2] - d$estimate[1]
```

```
mean of y
0.2947018
```

Der Unterschied beträg 0.29kg.

b) Berücksichtigen Sie nur die Daten der Mütter, die *während* der Schwangerschaft nicht geraucht haben. Berechnen Sie das 95%-Konfidenzintervall für den durchscnnittlichen Unterschied des Geburtsgewichts zwischen Kindern von Müttern, die *vor* der Schwangerschaft geraucht haben, und den Nichtraucherinnen.

```
# subsets
k  <- subset(neonates, smoke=="No")
k1 <- subset(k, smoke.before=="Yes")
k2 <- subset(k, smoke.before=="No")

# t-Test
d <- t.test(k1$weight, k2$weight, conf.level = 0.95)

# Konfidenzgrenzen
as.numeric(d$conf.int)
```

```
[1] -0.23226491  0.03252437
```

```
# Unterschied
d$estimate[2] - d$estimate[1]
```

```
 mean of y
0.09987027
```

Das Konfidenzintervall enthält die 0, das heisst, wir können nicht von einem signifikanten Unterschied ausgehen.

c) Berechnen Sie das 95%-Konfidenzintervall für den durchschnittlichen Unterschied von APGAR-1-Werten und APGAR-5-Werten. Wie entwickeln sich Neugeborene in den ersten 5 Minuten nach der Geburt?

```
# t-Test
d <- t.test(neonates$apgar1, neonates$apgar5, conf.level = 0.95)

# Konfidenzgrenzen
as.numeric(d$conf.int)
```

```
[1] -0.8093862 -0.3593638
```

```
# Unterschied
d$estimate[2] - d$estimate[1]
```

```
mean of y
 0.584375
```

Nach 5 Minuten haben die Neugeborenen einen im Schnitt 0,58 Punkte höheren APGAR-Wert.

d) Wenn Neugeborene mit einem APGAR-1-Wert ≤ 3 in einem kritischen Zustand sind, berechnen Sie das 90%-Konfidenzintervall für den Unterschied der Anteile von Neugeborenen in kritischem Zustand zwischen Müttern, die *während* der Schwangerschaft geraucht haben und den Nichtraucherinnen.

```
# neue Variable "kritisch"
neonates$kritisch <- "normal"
# nur solche mit APGAR<4 sind kritisch
neonates$kritisch[neonates$apgar1<4] <- "kritisch"

freq <- table(neonates$kritisch, neonates$smoke)

d <- prop.test(c(freq[["kritisch","No"]], freq[["kritisch","Yes"]]),
               c(sum(freq[,"No"]), sum(freq[,"Yes"])),
               alternative="two.sided", conf.level=0.90)



# Konfidenzgrenzen
as.numeric(d$conf.int)
```

```
[1] -0.22157738 -0.07478626
```

```
# Unterschied
d$estimate[2] - d$estimate[1]
```

```
   prop 2
0.1481818
```

Bei Raucherinnen sind durchschnittlich 15% mehr Neugeborene im kritischen Zustand zu finden als bei Nichtraucherinnen.

e) Hat das Alter der Mutter einen signifikanten Einfluss auf den Anteil an Neugeborenen in kritischem Zustand?

```
# neue Variable "kritisch"
neonates$kritisch <- "normal"
# nur solche mit APGAR<4 sind kritisch
neonates$kritisch[neonates$apgar1<4] <- "kritisch"

freq <- table(neonates$kritisch, neonates$age)

d <- prop.test(c(freq[["kritisch","greater than 20"]],
                 freq[["kritisch","less than 20"]]),
               c(sum(freq[,"greater than 20"]),
                 sum(freq[,"less than 20"])),
               alternative="two.sided", conf.level=0.95)



# Konfidenzgrenzen
as.numeric(d$conf.int)
```

```
[1] -0.151048570  0.006238353
```

```
# Unterschied
d$estimate[2] - d$estimate[1]
```

```
    prop 2
0.07240511
```

Das 95% Intervall enthält die 0. Wir können keinen signifikanten Unterschied feststellen.

2.63 Lösung zur Aufgabe 1.10.1

a) Übertragen Sie die Daten in ein Datenframe mit der Variable `Konzentration`.

```
Konzentration <- c(17.6, 19.2, 21.3, 15.1, 17.6, 18.9, 16.2, 18.3, 19.0, 16.4)
```

b) Testen Sie die zweiseitige Hypothese $H_0 : \mu = 18$ versus $H_1 : \mu \neq 18$ mit einem Signifikanzniveau von $\alpha = 0,05$.

```
t.test (Konzentration, alternative="two.sided", mu=18, conf.level=0.95)
```

```
    One Sample t-test

data:  Konzentration
t = -0.07078, df = 9, p-value = 0.9451
```

```
alternative hypothesis: true mean is not equal to 18
95 percent confidence interval:
 16.68158 19.23842
sample estimates:
mean of x
    17.96
```

Das Ergebnis ist nicht signifikant.

c) Testen Sie die zweiseitige Hypothese $H_0 : \mu = 19,5$ versus $H_1 : \mu \neq 19,5$ mit den Signifikanzniveaus von $\alpha = 0,05$ und $0,01$. Wie beeinflusst das Signifikanzniveau das Testergebnis?

```
t.test (Konzentration, alternative="two.sided", mu=19.5, conf.level=0.95)
```

```
    One Sample t-test

data:  Konzentration
t = -2.725, df = 9, p-value = 0.02341
alternative hypothesis: true mean is not equal to 19.5
95 percent confidence interval:
 16.68158 19.23842
sample estimates:
mean of x
    17.96
```

```
t.test (Konzentration, alternative="two.sided", mu=19.5, conf.level=0.99)
```

```
    One Sample t-test

data:  Konzentration
t = -2.725, df = 9, p-value = 0.02341
alternative hypothesis: true mean is not equal to 19.5
99 percent confidence interval:
 16.1234 19.7966
sample estimates:
mean of x
    17.96
```

Da der p-Wert bei $0,02341$ liegt, ist das Ergebnis für $\alpha = 0,05$ signifikant, für $\alpha = 0,01$ jedoch nicht.

d) Testen Sie die zweiseitige Hypothese $H_0 : \mu = 17$ versus $H_1 : \mu \neq 17$ mit einem Signifikanzniveau von $\alpha = 0,05$. Testen Sie ebenfalls die Hypothesen $H_0 : \mu = 17$ versus $H_1 : \mu > 17$ mit $\alpha = 0,05$. Was ist der Unterschied zwischen den p-Werten des zweiseitigen und des einseitigen Tests?

```
t.test (Konzentration, alternative="two.sided", mu=17, conf.level=0.95)
```

```
    One Sample t-test

data:  Konzentration
t = 1.6987, df = 9, p-value = 0.1236
alternative hypothesis: true mean is not equal to 17
95 percent confidence interval:
 16.68158 19.23842
sample estimates:
mean of x
    17.96
```

```
t.test (Konzentration, alternative="greater", mu=17, conf.level=0.95)
```

```
    One Sample t-test

data:  Konzentration
t = 1.6987, df = 9, p-value = 0.0618
alternative hypothesis: true mean is greater than 17
95 percent confidence interval:
 16.92404      Inf
sample estimates:
mean of x
    17.96
```

Der p-Wert ist beim einseitigen Test kleiner. Beide Werte sind jedoch größer als $0,05$.

e) Wenn der Hersteller angibt, die Konzentration des Wirkstoffs erhöht zu haben (im Vergleich zu früheren Chargen, bei denen der Mittelwert der Konzentration 17 mg/mm^3 war), können wir ihm glauben?

```
t.test (Konzentration, alternative="greater", mu=17, conf.level=0.95)
```

```
    One Sample t-test

data:  Konzentration
t = 1.6987, df = 9, p-value = 0.0618
alternative hypothesis: true mean is greater than 17
95 percent confidence interval:
 16.92404      Inf
sample estimates:
mean of x
```

```
    17.96
```

Der p-Wert ist nicht signifikant. Wir können dem Hersteller also nicht glauben.

f) Welche Fallzahl würde benötigt, um einen Konzentrationsanstieg von 0,5 mg/mm^3 zu erkennen (mit $\alpha = 0,05$ und einer Power von $1 - \beta = 0,8$)?

```
# Power-t-test
power.t.test(delta=0.5, sd=sd(Konzentration),
             sig.level=0.05, power=0.8, type = "one.sample")
```

```
     One-sample t test power calculation

              n = 102.2077
          delta = 0.5
             sd = 1.787114
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Es wird eine Fallzahl von 103 benötigt.

2.64 Lösung zur Aufgabe 1.10.2

a) Übertragen Sie die Daten in ein Datenframe mit der Variable `bib`.

```
# Daten übertragen
bib <- c("nein",  "ja",  "nein",  "nein",  "nein",  "ja",  "nein",
         "ja", "ja",  "ja",  "ja",  "nein",  "ja",  "nein", "ja",
         "nein",  "nein", "nein",  "ja",  "ja",  "ja",  "nein",
         "nein",  "ja",  "nein",  "nein", "ja",  "ja",  "nein",
         "nein",  "ja",  "nein",  "ja", "nein")
```

b) Testen Sie die Hypothese, dass der Anteil an Studierenden, die wöchentlich die Bibliothek nutzen, größer als 40% ist.

```
freq <- table(bib)
# testen
prop.test(freq[["ja"]], sum(freq), alternative="greater", p=0.4, conf.level=0.95)
```

```
    1-sample proportions test with continuity correction

data:  freq[["ja"]] out of sum(freq), null probability 0.4
X-squared = 0.4424, df = 1, p-value = 0.253
alternative hypothesis: true p is greater than 0.4
```

```
95 percent confidence interval:
 0.3238772 1.0000000
sample estimates:
        p
0.4705882
```

Der Test ist nicht signifikant.

2.65 Lösung zur Aufgabe 1.10.3

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `Alter` und `Population`.

```
# Daten übertragen
df <- data.frame(Alter = c(9.5, 10.5, 9.0, 9.8, 10.0, 13.0,
                           10.0, 13.5, 10.0, 9.8, 12.5, 9.5,
                           13.5, 13.8, 12.0, 13.8, 12.5, 9.5,
                           12.0, 13.5, 12.0, 12.0),
                 Population = c(rep("A", 10), rep("B", 12)))
```

b) Testen Sie die Hypothese, dass das durchschnittliche Alter in den Populationen unterschiedlich ist, mit $\alpha = 0,05$.

```
# teste, ob Varianzhomogenität vorliegt
var.test(Alter ~ Population, data=df)$p.value
```

```
[1] 0.9164489
```

```
# liegt vor
t.test(Alter ~ Population, data=df, var.equal=TRUE)
```

```
    Two Sample t-test

data:  Alter by Population
t = -2.6982, df = 20, p-value = 0.01383
alternative hypothesis: true difference in means between group A and group B is not equal to
95 percent confidence interval:
 -3.0260864 -0.3872469
sample estimates:
mean in group A mean in group B
       10.51000        12.21667
```

Das Ergebnis ist signifikant, p ist kleiner als 0,05. Es liegt also ein Unterschied vor.

2.66 Lösung zur Aufgabe 1.10.4

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `vorher` und `nachher`.

```
# Daten übertragen
df <- data.frame(vorher = c(60.6, 12.0, 56.0, 75.2, 12.5, 29.7,
                            57.2, 62.7, 28.7, 66.0, 25.2, 40.1),
                 nachher = c(47.5, 13.3, 33.0, 55.2, 21.9, 27.9,
                             54.3, 13.9, 8.90, 46.1, 29.8, 36.2))
```

b) Testen Sie, ob sich die Bronchialretention nach dem Rauchstopp verringert.

```
# Daten übertragen
t.test (df$vorher, df$nachher, alternative="greater", paired=TRUE, conf.level=0.95)
```

```
    Paired t-test

data:  df$vorher and df$nachher
t = 2.4847, df = 11, p-value = 0.01516
alternative hypothesis: true mean difference is greater than 0
95 percent confidence interval:
 3.185837      Inf
sample estimates:
mean difference
       11.49167
```

Das Ergebnis ist signifikant, p ist kleiner als 0,05. Es liegt also ein Unterschied vor, die Retention hat sich verringert.

2.67 Lösung zur Aufgabe 1.10.5

Gibt es signifikante Unterschiede zwischen den Prozentsätzen der Studiereden, die am Vormittag und am Nachmittag bestanden haben? Kann man daraus schließen, dass der Stundenplan die Ursache für diese Unterschiede ist?

```
# Daten übertragen
df <- data.frame(course = c(rep("bestanden", 55), rep("durchgefallen", 25),
                            rep("bestanden", 32), rep("durchgefallen", 58)),
                 time = c(rep("morgens", 80), rep("abends", 90))

                 )

freq <- table(df)
prop.test(c(freq[["bestanden","morgens"]], freq[["bestanden","abends"]]),
          c(sum(freq[,"morgens"]), sum(freq[,"abends"])),
          alternative="two.sided", conf.level=0.95)
```

```
    2-sample test for equality of proportions with continuity correction

data:  c(freq[["bestanden", "morgens"]], freq[["bestanden", "abends"]]) out of c(sum(freq[,
X-squared = 17.372, df = 1, p-value = 3.072e-05
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1783764 0.4855125
sample estimates:
   prop 1    prop 2
0.6875000 0.3555556
```

Das Ergebnis ist signifikant, p ist kleiner als 0,05. Es liegt also ein Unterschied zwischen morgens und abends vor.

2.68 Lösung zur Aufgabe 1.10.6

```
# lade Datensatz
load(url("https://www.produnis.de/R/data/pulse.RData"))
```

a) Testen Sie, ob der Ruhepuls weniger als 75 Schläge pro Minute beträgt.

```
t.test(pulse$pulse1, mu=75, alternative = "less")
```

```
    One Sample t-test

data:  pulse$pulse1
t = -1.8562, df = 91, p-value = 0.03333
alternative hypothesis: true mean is less than 75
95 percent confidence interval:
     -Inf 74.77684
sample estimates:
mean of x
 72.86957
```

Das Ergebnis ist signifikant.

b) Welcher Stichprobenumfang ist erforderlich, um einen Anstieg des Ruhepulses um 2 Schläge pro Minute mit einem Signifikanzniveau von 0,05 und einer Power von 0,9 festzustellen?

```
power.t.test(delta=2, sd=sd(pulse$pulse1),
             sig.level=0.05, power=0.9)
```

```
     Two-sample t test power calculation

              n = 637.6676
          delta = 2
```

```
             sd = 11.00871
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

Es werden 638 Probanden benötigt.

c) Testen Sie, ob der Puls nach dem Laufen größer als 85 Schläge pro Minute ist.

```
t.test(pulse$pulse2, mu=85, alternative="greater")
```

```
    One Sample t-test

data:  pulse$pulse2
t = -2.8056, df = 91, p-value = 0.9969
alternative hypothesis: true mean is greater than 85
95 percent confidence interval:
 77.03847      Inf
sample estimates:
mean of x
       80
```

Das Ergebnis ist nicht signifikant

d) Eine Person hat eine leichte Tachykardie, wenn der Ruhepuls größer als 90 Schläge pro Minute ist. Prüfen Sie, ob der Prozentsatz der Personen mit leichter Tachykardie größer als 5% ist.

```
pulse$tachy <- "nein"
pulse$tachy[pulse$pulse1 > 90] <- "ja"

freq <- table(pulse$tachy)
prop.test(freq[["ja"]], sum(freq), alternative="greater",
          p=0.05, conf.level=0.95)
```

```
    1-sample proportions test with continuity correction

data:  freq[["ja"]] out of sum(freq), null probability 0.05
X-squared = 0.18535, df = 1, p-value = 0.3334
alternative hypothesis: true p is greater than 0.05
95 percent confidence interval:
 0.03035962 1.00000000
sample estimates:
         p
0.06521739
```

Das Ergebnis ist nicht signifikant.

e) Kann man mit 95%iger Sicherheit schließen, dass Bewegung den Puls erhöht? Und bei einem Signifikanzniveau von $\alpha = 0,01$?

```
# test ob pulse1 kleiner ist als pulse2
t.test(pulse$pulse1, pulse$pulse2, alternative="less", conf.level = 0.95)
```

```
    Welch Two Sample t-test

data:  pulse$pulse1 and pulse$pulse2
t = -3.3638, df = 155.41, p-value = 0.0004841
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -3.622838
sample estimates:
mean of x mean of y
 72.86957  80.00000
```

```
t.test(pulse$pulse1, pulse$pulse2, alternative="less", conf.level = 0.99)
```

```
    Welch Two Sample t-test

data:  pulse$pulse1 and pulse$pulse2
t = -3.3638, df = 155.41, p-value = 0.0004841
alternative hypothesis: true difference in means is less than 0
99 percent confidence interval:
      -Inf -2.147776
sample estimates:
mean of x mean of y
 72.86957  80.00000
```

Das Ergebnis ist in beiden Fällen signifikant. Bewegung erhöht also den Puls.

f) Gibt es einen Unterschied zwischen den durchschnittlichen Pulsschlägen nach dem Gehen und dem Laufen?

```
# test ob pulse1 kleiner ist als pulse2
t.test(pulse2 ~ type, data=pulse, conf.level = 0.95)
```

```
    Welch Two Sample t-test

data:  pulse2 by type
t = 5.8335, df = 45.695, p-value = 5.251e-07
alternative hypothesis: true difference in means between group running and group walking is
95 percent confidence interval:
 13.22755 27.16944
```

```
sample estimates:
mean in group running mean in group walking
              92.51429               72.31579
```

Es gibt einen signifikanten Unterschied.

g) Gibt es einen Unterschied zwischen den Mittelwerten des Ruhepulses von Männern und Frauen? Und nach dem Laufen?

```
# test ob pulse1 kleiner ist als pulse2
t.test(pulse1 ~ sex, data=pulse, conf.level = 0.95)
```

```
    Welch Two Sample t-test

data:  pulse1 by sex
t = -2.7217, df = 63.675, p-value = 0.008367
alternative hypothesis: true difference in means between group male and group female is not
95 percent confidence interval:
 -11.160619  -1.711561
sample estimates:
  mean in group male mean in group female
            70.42105              76.85714
```

```
t.test(pulse2 ~ sex, data=pulse, conf.level = 0.95)
```

```
    Welch Two Sample t-test

data:  pulse2 by sex
t = -2.7849, df = 51.047, p-value = 0.007494
alternative hypothesis: true difference in means between group male and group female is not
95 percent confidence interval:
 -18.64912  -3.02507
sample estimates:
  mean in group male mean in group female
            75.87719              86.71429
```

Für beide Pulse gibt es signifikante Unterschiede zwischen Männern und Frauen.

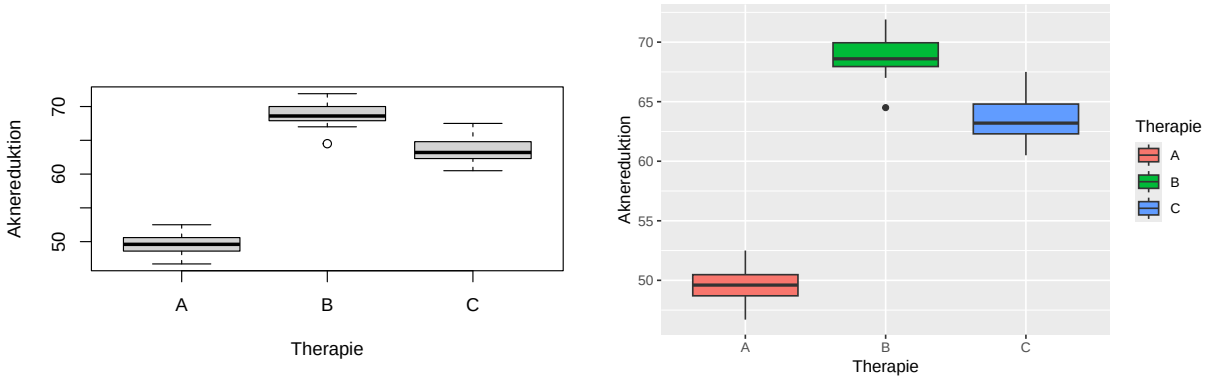
2.69 Lösung zur Aufgabe 1.11.1

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `Therapie` und `Aknereduktion`.

```
A <- c(48.6, 50.8, 49.4, 47.1, 50.1, 52.5, 49.8, 49, 50.6, 46.7)
B <- c(68, 71.9, 67, 71.5, 70.1, 69.9, 64.5, 68.9, 68, 67.8, 68.3, 68.9)
C <- c(67.5, 61.4, 62.5, 67.4, 64.2, 65.4, 62.5, 63.2, 63.9, 61.2, 64.8, 60.5, 62.3)

df <- data.frame(Therapie = c(rep("A", length(A)),
                              rep("B", length(B)),
                              rep("C", length(C))),
                 Aknereduktion = c(A,B,C)
)
```

b) Plotten Sie die Aknereduktion für jede Therapie. Sind Unterschiede erkennbar?

```
# plot()
boxplot(Aknereduktion ~ Therapie, data=df)
# ggplot()
ggplot(df, aes(x=Therapie, y=Aknereduktion)) +
  geom_boxplot(aes(fill=Therapie))
```



Therapie A unterscheidet sich deutlich von B und C.

c) Führen Sie eine ANOVA durch. Gibt es signifikante Unterschiede zwischen den Therapien?

```
fit <- aov(Aknereduktion ~ Therapie, data=df)
summary(fit)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
Therapie     2 2133.7  1066.8     262 <2e-16 ***
Residuals   32  130.3     4.1
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Es gibt Unterschiede. Schauen wir genauer hin.

```
pairwise.t.test(df$Aknereduktion, df$Therapie, p.adjust="bonferroni")
```

```
    Pairwise comparisons using t tests with pooled SD

data:  df$Aknereduktion and df$Therapie

  A       B
B < 2e-16 -
C < 2e-16 1.2e-06

P value adjustment method: bonferroni
```

Alle Gruppen unterscheiden sich jeweils voneinander.

d) Berechnen Sie die Konfidenzintervalle für die paarweisen Unterschiede zwischen den drei Behandlungen. Bei welchen Behandlungen gibt es signifikante Unterschiede?

```
ab <- t.test(A,B)
ac <- t.test(A,C)
cb <- t.test(C,B)

# Konfidenzintervalle
as.numeric(ab$conf.int)
```

```
[1] -20.93395 -17.61271
```

```
as.numeric(ac$conf.int)
```

```
[1] -15.85486 -12.42514
```

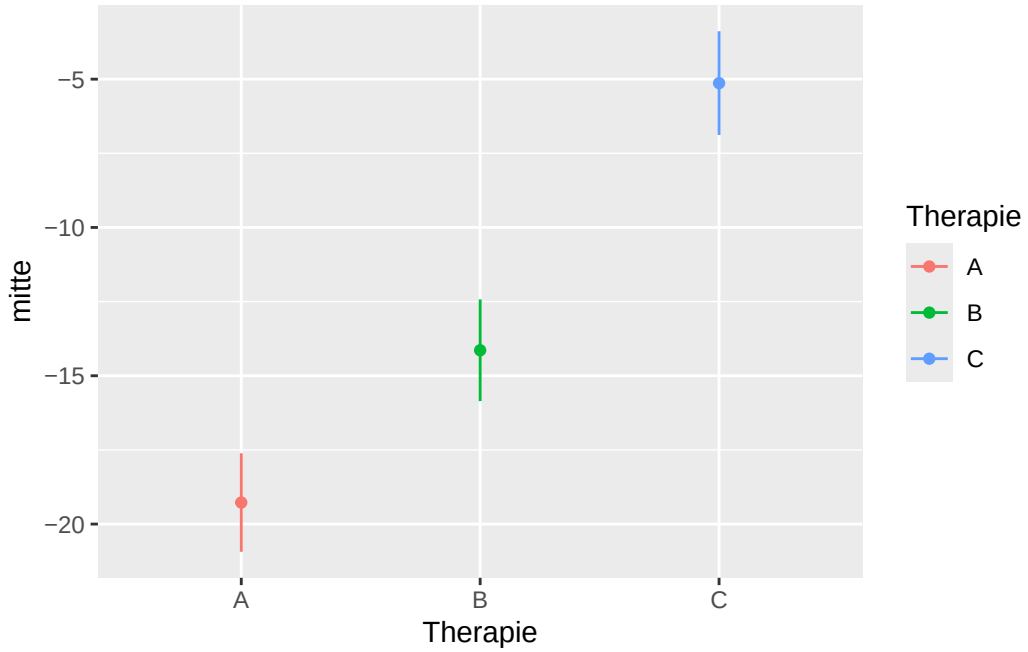
```
as.numeric(cb$conf.int)
```

```
[1] -6.880637 -3.386029
```

Keines der Intervalle schließt die 0 ein. Alle Therapien unterscheiden sich jeweils voneinander.

e) Plotten Sie diese Konfidenzintervalle.

```
# Als Datenframe für ggplot zusammenbauen
plotdf <- data.frame(unten = c(ab$conf.int[1], ac$conf.int[1], cb$conf.int[1]),
                     oben  = c(ab$conf.int[2], ac$conf.int[2], cb$conf.int[2]),
                     mitte = c(ab$estimate[1]-ab$estimate[2],
                               ac$estimate[1]-ac$estimate[2],
                               cb$estimate[1]-cb$estimate[2] ),
                     Therapie = c("A", "B", "C"))

# ggplot
ggplot(plotdf, aes(x=Therapie, color=Therapie)) +
  geom_point(aes(y=mitte)) +
  geom_segment(aes(xend=Therapie, y=unten, yend=oben))
```



Keines der Intervalle schließt die 0 ein. Alle Therapien unterscheiden sich jeweils voneinander.

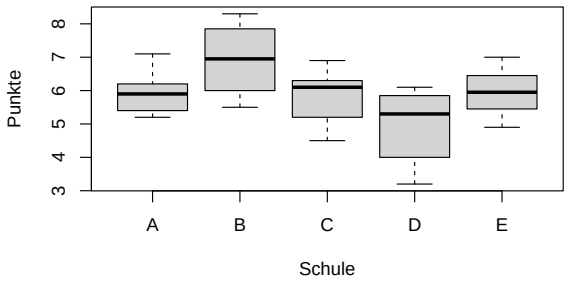
2.70 Lösung zur Aufgabe 1.11.2

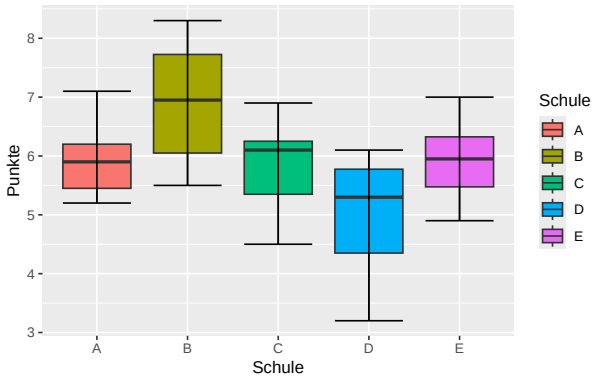
a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `Schule` und `Punkte`.

```
# Erstellen des tibbles mit tribble()
df <- tribble(
  ~A,   ~B,   ~C,   ~D,   ~E,
  5.5,  6.1,  4.9,  3.2,  6.7,
  5.2,  7.2,  5.5,  3.3,  5.8,
  5.9,  5.5,  6.1,  5.5,  5.4,
  7.1,  6.7,  6.1,  5.7,  5.5,
  6.2,  7.6,  6.2,  6.0,  4.9,
  5.9,  5.9,  6.4,  6.1,  6.2,
  5.3,  8.1,  6.9,  4.7,  6.1,
  6.2,  8.3,  4.5,  5.1,  7.0
) %>% # und pivot_longer
  pivot_longer(A:E, names_to="Schule", values_to="Punkte")
```

b) Plotten Sie die durchschnittlich erreichten Punkte pro Schule. Sind Unterschiede erkennbar?

```
# plot()
boxplot(Punkte ~ Schule, data=df)
# ggplot()
ggplot(df, aes(x=Schule, fill=Schule, y=Punkte)) +
  geom_boxplot() +
  # whiskers
  stat_boxplot(geom="errorbar")
```





c) Führen Sie eine ANOVA durch. Gibt es signifikante Unterschiede zwischen den Schulen?

```
fit <- aov(Punkte ~ Schule, data=df)
summary(fit)
```

```
             Df Sum Sq Mean Sq F value  Pr(>F)
```

```
Schule       4  15.69   3.921   5.031 0.00261 **
Residuals   35  27.28   0.779
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d) In welcher Schule sind die sportlichen Leistungen am besten?

```
pairwise.t.test(df$Punkte, df$Schule, p.adjust="bonferroni")
```

```
    Pairwise comparisons using t tests with pooled SD

data:  df$Punkte and df$Schule

  A       B       C       D
B 0.27923 -       -       -
C 1.00000 0.17589 -       -
D 0.36032 0.00078 0.55345 -
E 1.00000 0.33829 1.00000 0.29780

P value adjustment method: bonferroni
```

Es gibt einen Unterschied zwischen Schule B und Schule D. In Schule B sind die Leistungen besser als in Schule D. Die Leistungen in B sind aber nicht “die besten”, da kein Unterschied zu den anderen Schulen gezeigt werden kann. Die Mittelwerte sind in B aber höher.

2.71 Lösung zur Aufgabe 1.11.3

Gibt es laut den Daten signifikante Unterschiede zwischen den vier Gruppen?

```
# Erstellen des tibbles mit tribble()
df <- tribble(
  ~Kontrolle,  ~AnginaP,  ~Arrhytmia,  ~Herzinfarkt,
          83,  81,  75,  61,
          61,  65,  68,  75,
          80,  77,  80,  78,
          63,  87,  80,  80,
          67,  95,  74,  68,
          89,  89,  78,  65,
          71,  103, 69,  68,
          73,  89,  72,  69,
          70,  78,  76,  70,
          66,  83,  75,  79,
          57,  91,  69,  61
) %>% # und pivot_longer
  pivot_longer(1:4, names_to="Gruppe", values_to="Puls")

# ANOVa
fit <- aov(Puls ~ Gruppe, data=df)
summary(fit)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
Gruppe       3   1587   529.1   8.043 0.00026 ***
Residuals   40   2631    65.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# paarweise
pairwise.t.test(df$Puls, df$Gruppe, p.adjust="bonferroni")
```

```
    Pairwise comparisons using t tests with pooled SD

data:  df$Puls and df$Gruppe

            AnginaP Arrhytmia Herzinfarkt
Arrhytmia   0.01584 -         -
Herzinfarkt 0.00062 1.00000   -
Kontrolle   0.00100 1.00000   1.00000

P value adjustment method: bonferroni
```

Patienten mit Angina Pectoris unterscheiden sich von allen anderen Patientengruppen.

2.72 Lösung zur Aufgabe 1.11.4

Gibt es laut den Daten signifikante Unterschiede zwischen den drei Gruppen?

```
# Erstellen des tibbles mit tribble()
df <- tribble(
  ~Low, ~Medium, ~High,
  36,    43,     45,
  33,    38,     39,
  35,    41,     33,
  39,    34,     39,
  41,    28,     33,
  41,    44,     26,
  44,    30,     39,
  45,    31,     29
) %>% # und pivot_longer
  pivot_longer(1:3, names_to="Konzentration", values_to="Atemfrequenz")

# ANOVa
fit <- aov(Atemfrequenz ~ Konzentration, data=df)
summary(fit)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Konzentration  2   67.6   33.79   1.056  0.366
Residuals     21  672.2   32.01
```

```
# paarweise
pairwise.t.test(df$Atemfrequenz, df$Konzentration, p.adjust="bonferroni")
```

```
    Pairwise comparisons using t tests with pooled SD

data:  df$Atemfrequenz and df$Konzentration

       High Low
Low    0.56 -
Medium 1.00 0.85

P value adjustment method: bonferroni
```

Es kann kein signifikanter Unterschied festgestellt werden.

2.73 Lösung zur Aufgabe 1.12.1

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `Geschwuer` und `Blutgruppe`.

```
df <- data.frame(Geschwuer=c(rep("Geschwür", 1655),
                             rep("gesund", 10000)),
                 Blutgruppe=c(rep("0", 911), rep("A", 579),
                              rep("B", 124), rep("AB", 41),
                              rep("0", 4578), rep("A", 4219),
                              rep("B", 890), rep("AB", 313))
)
```

b) Führen Sie einen Chiquadrattest auf die Hypothese durch, dass die Geschwüre von der Blutgruppe abhängig sind.

```
chisq.test(df$Geschwuer, df$Blutgruppe)
```

```
    Pearson's Chi-squared test

data:  df$Geschwuer and df$Blutgruppe
X-squared = 49.016, df = 3, p-value = 1.295e-10
```

Es gibt unterschiede, p ist kleiner als 0,05.

c) Gibt es in Anbetracht der Ergebnisse des Vergleichs einen Zusammenhang zwischen dem Magengeschwür und der Blutgruppe? Können wir behaupten, dass der Anteil der Ulkuspatienten je nach Blutgruppe unterschiedlich ist?

```
reporttools::pairwise.fisher.test(df$Geschwuer, df$Blutgruppe,
                                  p.adjust.method = "bonferroni")
```

```
    Pairwise comparisons using

data:  df$Geschwuer and df$Blutgruppe

   0       A       AB
A  4.1e-10 -       -
AB 0.0695  1.0000  -
B  0.0023  1.0000  1.0000

P value adjustment method: bonferroni
```

Blutgruppe 0 unterscheidet sich von allen anderen.

2.74 Lösung zur Aufgabe 1.12.2

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `Region` und `Blutgruppe`.

```
df <- data.frame(
          Blutgruppe=c(rep("A", 185), rep("B", 55),
                       rep("0", 223), rep("AB", 15)),
          Region=c(rep("Eskdale", 33), rep("Annadale", 54), rep("Nithsdale", 98),
                   rep("Eskdale", 6), rep("Annadale", 14), rep("Nithsdale", 35),
                   rep("Eskdale", 56), rep("Annadale", 52), rep("Nithsdale", 115),
                   rep("Eskdale", 5), rep("Annadale", 5), rep("Nithsdale", 5))
)
```

b) Führen Sie einen Chiquadrattest auf die Hypothese durch, dass die Blutgruppe von der Region abhängig sind.

```
chisq.test(df$Blutgruppe, df$Region)
```

```
Warning in chisq.test(df$Blutgruppe, df$Region): Chi-Quadrat-Approximation kann
inkorrekt sein

    Pearson's Chi-squared test

data:  df$Blutgruppe and df$Region
X-squared = 10.454, df = 6, p-value = 0.1068
```

Der Test ist nicht signifikant.

c) Gibt es in Anbetracht der Ergebnisse einen Zusammenhang zwischen der Blutgruppe und der Region? Können wir behaupten, dass die Region keinen Einfluss auf die Blutgruppe hat?

```
reporttools::pairwise.fisher.test(df$Blutgruppe, df$Region,
                                  p.adjust.method = "bonferroni")
```

```
    Pairwise comparisons using

data:  df$Blutgruppe and df$Region

          Annadale Eskdale
Eskdale   0.39     -
Nithsdale 1.00     0.10

P value adjustment method: bonferroni
```

Es sind keine Unterschiede zu finden.

2.75 Lösung zur Aufgabe 1.12.3

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `Rauchen` und `Geschlecht`.

```
df <- data.frame(
          Geschlecht=c(rep("m", 9), rep("w", 17)),
          Rauchen=c(rep("ja", 2), rep("nein", 7),
                    rep("ja", 6), rep("nein", 11))
)
```

b) Führen Sie einen Chi-Quadrat-Test durch, um festzustellen, ob das Rauchen mit dem Geschlecht zusammenhängt.

```
chisq.test(df$Rauchen, df$Geschlecht)
```

```
Warning in chisq.test(df$Rauchen, df$Geschlecht): Chi-Quadrat-Approximation
kann inkorrekt sein

    Pearson's Chi-squared test with Yates' continuity correction

data:  df$Rauchen and df$Geschlecht
X-squared = 0.057825, df = 1, p-value = 0.81
```

```
# kleines sample, besser exakten Fisher Test
fisher.test(df$Rauchen, df$Geschlecht)
```

```
    Fisher's Exact Test for Count Data

data:  df$Rauchen and df$Geschlecht
p-value = 0.6673
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.04160546 4.26600654
sample estimates:
odds ratio
  0.536523
```

c) Ist die Verteilung der Raucher bei beiden Geschlechtern gleich?

```
# kleines sample, besser exakten Fisher Test
fisher.test(df$Rauchen, df$Geschlecht)
```

```
    Fisher's Exact Test for Count Data

data:  df$Rauchen and df$Geschlecht
p-value = 0.6673
alternative hypothesis: true odds ratio is not equal to 1
```

```
95 percent confidence interval:
 0.04160546 4.26600654
sample estimates:
odds ratio
  0.536523
```

Beide Tests sind nicht signifikant, es besteht kein Unterschied in den Gruppen.

2.76 Lösung zur Aufgabe 1.12.4

a) Übertragen Sie die Daten in ein Datenframe mit den Variablen `drug1` und `drug2`.

```
df <- data.frame(
      drug1 = c( "Ja", "Ja", "Ja", "Ja", "Ja", "Nein", "Ja", "Nein",
                 "Ja", "Ja", "Ja", "Nein", "Ja", "Nein", "Ja", "Ja",
                 "Ja", "Nein", "Ja", "Ja"),
      drug2 = c("Nein", "Nein", "Ja", "Nein", "Ja", "Ja", "Nein", "Nein",
                "Nein", "Nein", "Ja", "Nein", "Ja", "Nein", "Nein", "Ja",
                "Nein", "Ja", "Nein", "Nein")
)
```

b) Führen Sie einen McNemar-Test durch, um festzustellen, ob die Linderung mit dem Medikament zusammenhängt.

```
mcnemar.test(df$drug1, df$drug2)
```

```
    McNemar's Chi-squared test with continuity correction

data:  df$drug1 and df$drug2
McNemar's chi-squared = 4.0833, df = 1, p-value = 0.04331
```

Das Ergebnis ist signifikant. Es gibt einen Unterschied.

c) Können wir nach dem Ergebnis des Tests behaupten, dass die Linderung der Migräne vom Medikament abhängt? Wenn ja, welches Medikament bewirkt eine signifikant höhere Linderung?

```
prop.table(table(df$drug1))
```

```
  Ja Nein
0.75 0.25
```

```
prop.table(table(df$drug2))
```

```
  Ja Nein
0.35 0.65
```

Medikament `drug1` wirkt besser.

2.77 Lösung zur Aufgabe 1.12.5

Ist ein komatöser Zustand bei der Ankunft im Krankenhaus ein Risikofaktor zu versterben?

```
df <- data.frame(
      Ergebnis = c(rep("überlebt", 521), rep("verstorben", 207)),
      Komatös = c(rep("nein", 484), rep("ja", 37),
                  rep("nein", 118), rep("ja", 89)
  )
)

chisq.test(df$Ergebnis, df$Komatös)
```

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  df$Ergebnis and df$Komatös
X-squared = 130.86, df = 1, p-value < 2.2e-16
```

Ja, es gibt einen Unterschied zwischen komatösen und nicht-komatösen Patienten.

2.78 Lösung zur Aufgabe 1.12.6

Ist die Wirksamkeit der beiden Behandlungen die gleiche?

```
df <- data.frame(
      Therapie = c(rep("A", 32), rep("B", 28)),
      Wirkung = c(rep("Sehr gut", 10), rep("gut", 14), rep("schlecht", 8),
                  rep("Sehr gut", 12), rep("gut", 10), rep("schlecht", 6)
  )
)

chisq.test(df$Wirkung, df$Therapie)
```

```
    Pearson's Chi-squared test

data:  df$Wirkung and df$Therapie
X-squared = 0.87141, df = 2, p-value = 0.6468
```

Es kann kein signifikanter Unterschied gezeigt werden.

2.79 Lösung zur Aufgabe 1.12.7

Können wir dann behaupten, dass Frauen in diesem Fach erfolgreicher sind als Männer?

```
df <- data.frame(
      Geschlecht = c(rep("m", 10), rep("w", 10)),
      bestanden = c(rep("ja", 2), rep("nein", 8),
                    rep("ja", 4), rep("nein", 6)
  )
)

table(df)
```

```
          bestanden
Geschlecht ja nein
         m  2    8
         w  4    6
```

```
chisq.test(df$bestanden, df$Geschlecht)
```

```
Warning in chisq.test(df$bestanden, df$Geschlecht): Chi-Quadrat-Approximation
kann inkorrekt sein

    Pearson's Chi-squared test with Yates' continuity correction

data:  df$bestanden and df$Geschlecht
X-squared = 0.2381, df = 1, p-value = 0.6256
```

Es kann kein signifikanter Unterschied gezeigt werden.

2.80 Lösung zur Aufgabe 1.12.8

Können wir bestätigen, dass es unterschiedliche Meinungen über Hans und Erna gibt?

```
df <- data.frame(
      Erna = c(rep("ja", 85), rep("nein", 65)),
      Hans = c(rep("ja", 37), rep("nein", 48),
               rep("ja", 44), rep("nein", 21)
  )
)

table(df)
```

```
      Hans
Erna   ja nein
  ja   37   48
```

```
  nein 44   21
```

```
chisq.test(df$Erna, df$Hans)
```

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  df$Erna and df$Hans
X-squared = 7.712, df = 1, p-value = 0.005486
```

Es gibt einen Signifikanten Unterschied zwischen den Dozenten.

```
prop.table(table(df$Erna))
```

```
       ja      nein
0.5666667 0.4333333
```

```
prop.table(table(df$Hans))
```

```
  ja nein
0.54 0.46
```

Die Studierenden mögen Erna ein bisschen mehr als Hans.

Literatur

Gimeno, E. A., Garro, J. C., Alberca, A. S., & Zaragoza de Lorite, A. (2022). *Applied Biostatistics with R and rk.Teaching*. https://github.com/asalber/statistics_practice_rkteaching

große Schlarmann, J. (2024). *Statistik mit R und RStudio - Ein Nachschlagewerk für Gesundheitsberufe*. Hochschule Niederrhein. https://www.produnis.de/R

Mitchell, R. J., Izatt, M. M., Sunderland, E., & Cartwright, R. A. (1976). Blood groups antigens, plasma protein and red cell isoenzyme polymorphisms in South-west Scotland. *Annals of Human Biology*, *3*(2), 157–171. https://doi.org/10.1080/03014467600001271

Credits

Prof. Dr. Jörg große Schlarmann

Hochschule Niederrhein, Krefeld

joerg.grosseschlarmann@hs-niederrhein.de

https://www.produnis.de/R

https://github.com/produnis/angewandte_uebungen_in_R